NeOn: Lifecycle Support for Networked Ontologies

Integrated Project (IST-2005-027595)

Priority: IST-2004-2.4.7 – "Semantic-based knowledge and content systems"

7.3.1 : Results from Experiments in Ontology Learning including Evaluation and Recommendation

Deliverable Co-ordinator: Alfio Massimiliano Gliozzo

Deliverable Co-ordinating Institution: Italian National Research Council (CNR)

Other Authors: Caterina Caracciolo (FAO), Mathieu d'Aquin (OU), Marta Sabou (OU) , Wim Peters (USFD), Johanna Voelker (UKARL)

Document Identifier:	NEON/2007/D7.3.1/v7.2	Date due:	October 31, 2007
Class Deliverable:	NEON EU-IST-2005-027595	Submission date:	December 15, 2007
Project start date:	March 1, 2006	Version:	V7.2
Project duration:	4 years	State:	Final
		Distribution:	Restricted

NeOn Consortium

This document is a part of the NeOn research project funded by the IST Programme of the Commission of the European Communities by the grant number IST-2005-027595. The following partners are involved in the project:

Open University (OU) – Coordinator Knowledge Media Institute – KMi Berrill Building, Walton Hall Milton Keynes, MK7 6AA United Kingdom Contact person: Martin Dzbor, Enrico Motta E-mail address: {m.dzbor, e.motta} @open.ac.uk	**Universität Karlsruhe – TH (UKARL)** Institut für Angewandte Informatik und Formale Beschreibungsverfahren – AIFB Englerstrasse 28 D-76128 Karlsruhe, Germany Contact person: Peter Haase E-mail address: pha@aifb.uni-karlsruhe.de
Universidad Politécnica de Madrid (UPM) Campus de Montegancedo 28660 Boadilla del Monte Spain Contact person: Asunción Gómez Pérez E-mail address: asun@fi.upm.es	**Software AG (SAG)** Uhlandstrasse 12 64297 Darmstadt Germany Contact person: Walter Waterfeld E-mail address: walter.waterfeld@softwareag.com
Intelligent Software Components S.A. (ISOCO) Calle de Pedro de Valdivia 10 28006 Madrid Spain Contact person: Jesús Contreras E-mail address: jcontreras@isoco.com	**Institut 'Jožef Stefan' (JSI)** Jamova 39 SI-1000 Ljubljana Slovenia Contact person: Marko Grobelnik E-mail address: marko.grobelnik@ijs.si
Institut National de Recherche en Informatique et en Automatique (INRIA) ZIRST – 655 avenue de l'Europe Montbonnot Saint Martin 38334 Saint-Ismier France Contact person: Jérôme Euzenat E-mail address: jerome.euzenat@inrialpes.fr	**University of Sheffield (USFD)** Dept. of Computer Science Regent Court 211 Portobello street S14DP Sheffield United Kingdom Contact person: Hamish Cunningham E-mail address: hamish@dcs.shef.ac.uk
Universität Koblenz-Landau (UKO-LD) Universitätsstrasse 1 56070 Koblenz Germany Contact person: Steffen Staab E-mail address: staab@uni-koblenz.de	**Consiglio Nazionale delle Ricerche (CNR)** Institute of cognitive sciences and technologies Via S. Martino della Battaglia, 44 - 00185 Roma-Lazio, Italy Contact person: Aldo Gangemi E-mail address: aldo.gangemi@istc.cnr.it
Ontoprise GmbH. (ONTO) Amalienbadstr. 36 (Raumfabrik 29) 76227 Karlsruhe Germany Contact person: Jürgen Angele E-mail address: angele@ontoprise.de	**Food and Agriculture Organization of the United Nations (FAO)** Viale delle Terme di Caracalla 1 00100 Rome Italy Contact person: Marta Iglesias E-mail address: marta.iglesias@fao.org
Atos Origin S.A. (ATOS) Calle de Albarracín, 25 28037 Madrid Spain Contact person: Tomás Pariente Lobo E-mail address: tomas.parientelobo@atosorigin.com	**Laboratorios KIN, S.A. (KIN)** C/Ciudad de Granada, 123 08018 Barcelona Spain Contact person: Antonio López E-mail address: alopez@kin.es

Work package participants

The following partners have taken an active part in the work leading to the elaboration of this document, even if they might not have directly contributed writing parts of this document: CNR, FAO, OU, USFD, UKARL

Change Log

Version	Date	Amended by	Changes
0.1	22/07/2007	Alfio Gliozzo (CNR)	Document Structure
0.2	29/08/2007	Alfio Gliozzo (CNR)	Editing content from the WIKI
1.0	01/10/2007	Alfio Gliozzo(CNR)	Contributions from USFD, section 4
1.1	16/10/2007	Alfio Gliozzo (CNR)	State of the art RE, Evaluaton LSA RE
1.2	25/10/2007	Alfio Gliozzo(CNR)	Contributions from OU, UKARL
1.3	27/10/2007	Alfio Gliozzo(CNR)	Minor syntactic revisions
2.0	30/10/2007	Alfio Gliozzo(CNR)	All contributions have been included and revised
2.1	31/10/2007	Caterina Caracciolo	Executive summary and Introduction amended
2.2	2/11/2007	Wim Peters	Amendment of section 3
3.0	3/11/2007	Alfio Gliozzo	Similarity induction, recommendation, bibliography, general revision of the document
4.0	5/11/2007	Alfio Gliozzo	Evaluation of terminology induction, recommendations.
4.1	14/11/2007	Caterina Caracciolo	Sec. 1.1.4, 1.1.5, Chapter 6 amended.
5.0	16/11/2007	Alfio Gliozzo	Titles of section 3 and 4, minor comments from Caterina. This is the version sent to QA
6.0	01/12/2007	Alfio Gliozzo	Adaptation to the QA requirements: revised conclusion (no rigid lower bound on accuracy), spelling, formatting issues, original smooth introduction reintegrated, Johanna's contributions to conclusion integrated
7.0	04/12/2007	Wim Peters	Style and spelling check
7.1	12/12/2007	Wim Peters	QA comments addressed
7.2	13/12/2007	Alfio Gliozzo	Final Revision and consistency check

Executive Summary

This document describes the Ontology Learning experiments we performed in order to recommend a set of ontology learning techniques to enhance the ontology engineering process in the fisheries domain in place at FAO. Experiments have been conceived in the wider context of the WorkPackage 7 of the NeOn Project. The main contribution of this document to WP7 is a set of recommendations and best practices to exploit semi-automatic technique to acquire knowledge either from domain specific documents and existing ontologies. The basic criterion for success indication adopted is the reduction of the development time required for the Ontology Engineering process actually in place at FAO. After a brief overview of WP7, the document addresses the following issues: to describe the state of the art in ontology learning and; to set up an evaluation case study in the fisheries domain; to identify suitable techniques which can be profitably applied to fit the user's requirements; to evaluate such techniques in the use case scenario, and, finally, to provide a set of recommendations indicating the most reliable techniques to be included in the ontology engineering lifecycle.

Table of Contents

List of tables

List of figures

1 Introduction

This document describes the work of task 7.3, concerning the evaluation of existing ontology learning techniques in the use case provided by WP 7, i.e. management of the ontology lifecycle in the fisheries domain. The document describes the techniques adopted, motivating their selection from a wider pool of technologies described in the state of the art section, and the achieved results both in terms of accuracy of each technique and in terms of applicability to the use case provided by the ontology lifecycle management in the fisheries domain in place at the Food and Agriculture Organization (FAO) of the United Nations. The outcome of this deliverable is then a set of recommended Ontology Learning technologies to be further integrated in the NeOn toolkit architecture described in D7.4.1 to provide well assessed, experimented and cost effective solutions to the "Populate from text" use case (described in D7.4.1, Use Case10).

The goal of these experiments is to demonstrate that existing state of the art ontology learning technologies can be applied to the fisheries use case and that their application leads to a sensible human effort reduction, providing quite accurate results by applying computationally efficient instruments. To achieve this goal we first identified a set of user's requirements, based on the operations of enriching, populating and mapping fisheries ontologies currently adopted at FAO. Such operations are very tedious and cost intensive, since they require *ontology editors* with consolidated expertise in the domain of interest. During the population task, especially when a totally new ontology is being developed, the number of instances, concepts and relations to be included in the ontology is typically huge (e.g. in the use case provided by the fisheries domain it involves thousands of concepts). In addition, it contains domain specific terms, which typically refer to scientific / domain specific literature, requiring a very high knowledge of the field to be understood, retrieved and then conceptualized. In addition, existing ontologies sometimes need to be mapped, for example by discovering semantic relations among their concepts. When the dimension of the ontology becomes huge, the number of concepts to be checked to perform such operation becomes unmanageable by means of a standard manual inspection or a keyword based search, forcing us to adopt semi-automatic techniques for mapping. The aim of this document is to demonstrate that by combining intelligent techniques for text processing, knowledge acquisition, reasoning, and by exploiting the semantic WEB infrastructure such as SWOOGLE, the effort required during the editorial process will be highly reduced, both in terms of the overall time required and in terms of coverage/accuracy of the ontology produced so far.

For example, automatically extracting terminology from domain specific texts ranked according to their frequency will provide the ontology editor with a (generally) complete list of concepts/instances in the domain of interest (and possibly the references to the documents where they have been found in the domain specific corpus). The more relevant terms will be included in the top ranked part of the list, indicating the concepts which *should* be included in the ontology, while looking at the bottom part of the list the ontology editor could find enough material to enhance coverage. In such a way, the probability of missing relevant concepts is sensibly reduced while the human effort required (and therefore the development time) is minimized. Terminology induction is just the simplest, relatively consolidated ontology learning technique. In this document we will also describe the exploitation of much more elaborated best practices in ontology learning, including Named Entity Recognition, Relation Extraction, Synonymy Induction and Ontology Mapping. For both the relation extraction and the synonymy induction experiments we will rely on a domain restriction hypothesis (Gliozzo et al., 2007), claiming that the probability of finding semantic relations between term is directly related to their topical similarity, which in turn can be estimated by adopting totally unsupervised corpus based techniques such as Latent Semantic Analysis (Landauer et al., 1990). The ontology engineering process benefits a lot from adopting such a

restriction, since it provides a method to filter out irrelevant relations, terms and associations, and to suggest the ontology engineer a set of plausible concepts/instances during the population phase. For the ontology mapping experiments we adopted a technique based on external knowledge accessed by using SWOOGLE.

Even though the achieved results support the claim that a fully automatic process for ontology learning from texts is still a chimera, very effective and practical techniques are now available to achieve partial goals. Nonetheless, we performed punctual and well motivated experiments, which show the ease of applicability of the proposed methods, and suggest that the impact of adopting such techniques in a large scale ontology engineering process would be highly beneficial, allowing drastic reduction of time and human effort, avoiding subjective judgments and conceptual gaps, and therefore augmenting the overall quality of the produced ontology.

1.1 Fisheries ontologies lifecycle

The lifecycle for fisheries ontologies identified within WP7 (cf. deliverable D7.4.1 [D7.4.1]) consists in three main phases (see Figure 1 below, from left to right):

1. the ontology is created and populated following an iterative process,
2. the ontology enters the maintenance phase (a workflow may be applied),
3. the ontology is published.

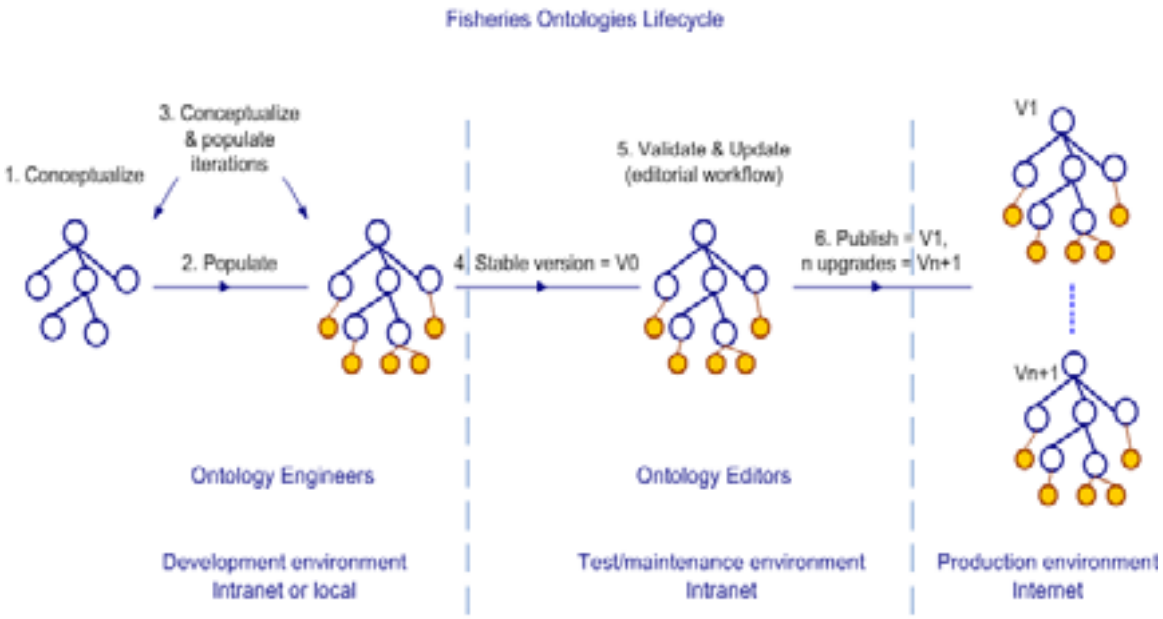

Figure 1. Main steps in the fisheries ontologies lifecycle.

The people playing a role in each phase are also to be distinguished: phase 1 is usually carried out by ontology engineers, phase 2 by ontology editors, while in phase 3 the users of the (published) ontologies include both programmers and casual visitors.

Task T7.3 mainly deals with the activities taking place in phase 2: the maintenance of the ontology carried on by ontology editors. Another use case for the experiments is the initial population of the ontology, taking place in phase 1 by ontology engineers.

When dealing with the maintenance of the ontology, the goal of the task is to investigate which techniques are suitable to support ontology editors in their daily work. The corresponding requirement can be found in D7.1.1 (Sec. 4.4.2 “Support to editing: ontology population”). Since editors should be supported in their daily work, not replaced, an appropriate degree of interaction with the user should be implemented. Given an ontology and one or more corpora, editors should receive “suggestions” for including concepts, instances, mapping or relations in the ontology they are working on. Each candidate is to be assessed by the editor and, if judged appropriate, included in the ontology (as “final” or “draft”, depending on the state in their workflow). Specific user requirements for the experiments carried out in this task are given in Section 1.2. Figure 2 depicts the possible ways in which ontologies can be populated in phase 1, when ontology engineers create ontologies following an iterative process of conceptualization and population. That figure (under the dashed green line) shows that there are at least three possible main sources for ontology population: manual creation, and import from a relational database or an XML schema. The fisheries ontologies currently available and described in deliverable D7.2.2 have been created from the fisheries reference data database and from XML schema for fisheries factsheets. In deliverable D7.2.2 we presented the lessons learned during those population activities. Also, in the case of manual population it can be useful to have some tools to support the ontology engineers in their activities: task T7.3 can then be applied also to this case.

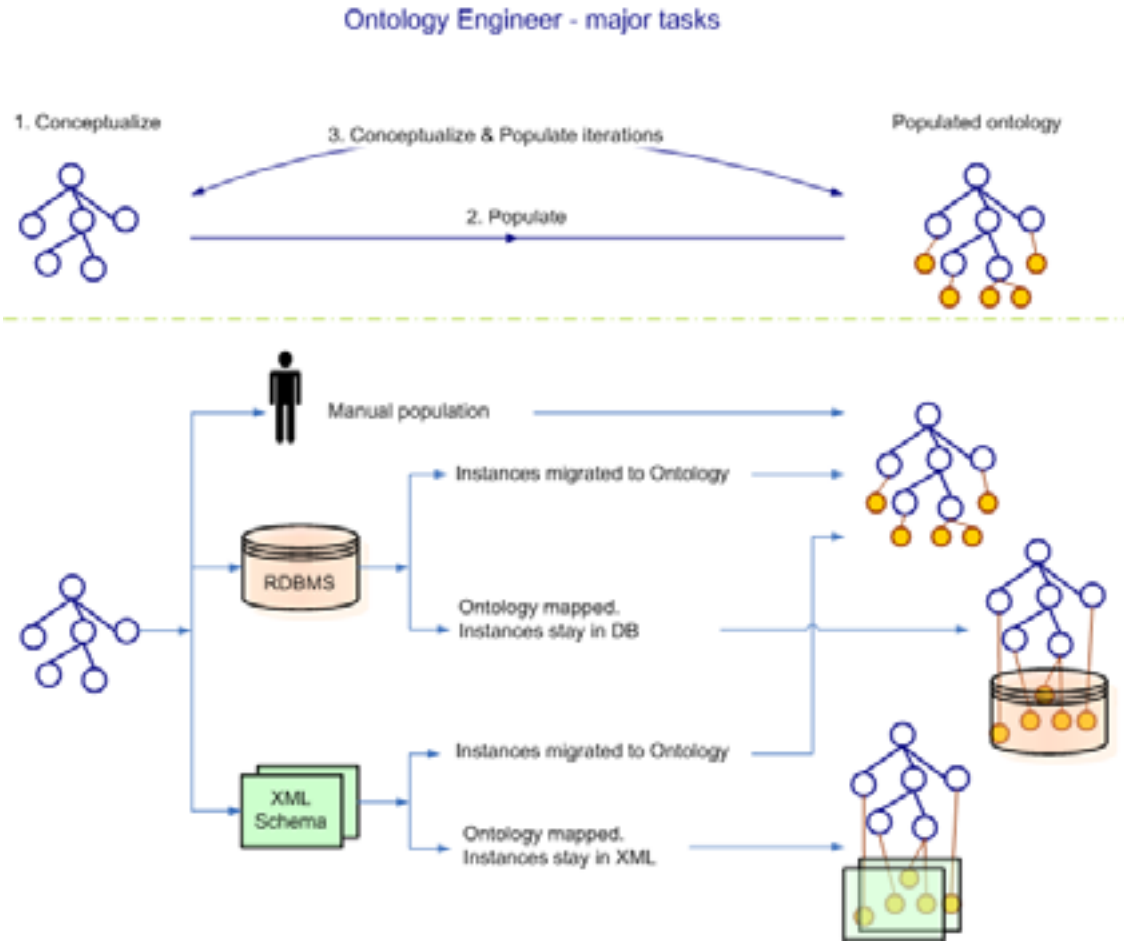

Figure 2. Methods for fishery ontologies population

In this deliverable we report on the work we carried out concerning the investigation and selection of suitable algorithms for future inclusion in the NeOn toolkit. The techniques adopted in our work are terminology extraction, similarity induction, relation extraction and ontology mapping.

The outcome of those experiments is an estimation of pros and cons related to the exploitation of such techniques rather than a set of effective tools to be directly integrated in the ontology engineering platform. Further steps of the project will concern the integration of the most successful techniques, identified in this Deliverable, in the overall Ontology Engineering lifecycle and their adaptation to different domains.

1.2 User Requirements

In this section we describe the user requirements elaborated for the experiment reported in Chapter 4. These user requirements integrate those given in D7.1.1, in particular in Sec. 7.4.2.

1.2.1 Users of the editing support tools

The expected users of the tools are the ontology editors (cf. D7.1.1). Ontology editors are either subject experts or information management specialists, who will be in charge of the everyday editing and maintenance work of the networked ontologies. Their duties also involve the creation of multilingual versions of ontologies.

Subject experts know about the domain to be modelled, but usually know little or nothing about ontology design issues or software for ontologies. Subject experts can be in charge of developing specific fragments of ontologies, revise work done by others, and develop multilingual versions of ontologies. Subject experts should be provided with more intuitive interfaces than those available to ontology experts and application developers; in particular interfaces for subject experts should conceal much of the purely ontological and engineering decisions.

Information management and technology (IM/IT) experts are specialists in the entire information management lifecycle, have some programming skills, or at least familiarity with implementation issues, and also have understanding of issues related to domain and domain modelling. IM/IT experts will often work on ontology maintenance, possibly in conjunction with subject experts.

1.2.2 Tasks to support

The experiments in T7.3 aim at selecting the best way to support the following types of editorial tasks:

1. **Ontology population**, to add appropriate classes or instances to the selected ontology (the modelling adopted in the ontology matters). One example for adding instances is to suggest alternative names of biological species starting from a document collection. Names can be in different languages, including scientific terminology based on Latin. Another example is to e.g. suggest fish species to add under a "family" (in a taxonomic or similar structure).
2. **Extracting relations between entities that can be stored in different ontologies**. For example, e.g. "Tuna fish" (as biological species) is the biological source of the commodity "canned tuna". The biological species X lives in water bodies Y, Z.
3. **Finding mappings between ontologies**. Mappings can link together entities that are equivalent or similar concepts in ASFA and AGROVOC (they overlap).

1.2.3 Data format

The experiments should use the most commonly used text file formats: PDF, HTML, DOC.

1.2.4 Human assessments

Experimental results are to be evaluated manually by domain experts in terminology or in fisheries related areas. Given their background, and the different nature of their daily activities, it is important that the activity of manual assessment be well focused. When a benchmark approach is not possible, the human evaluation should be organized in a way that:

1. domain experts are clearly introduced to the goals of the experiments and to the reasons why the specialist expertise is needed;
2. the actions necessary to perform the evaluation are clearly described by means of guidelines and instructions;
3. a graphical interface is available to the experts for evaluation.
    - The interface used for evaluation should be simple and intuitive in order to be adopted by domain experts in biology, fishery, oceanography and so on (as opposed to experts in computer science of ontology engineering).
    - Editors should always be shown the available pieces of information supporting the element to include:
        - a document excerpt,
        - document metadata (title of the document, author, data owner, publication date should be shown).

The graphical user interface will also be crucial if later inclusion in the NeOn toolkit is recommended, as domain experts should be facilitate in their work dealing with ontologies (cf. D7.1.1).

1.2.5 Evaluation of the results

The purpose of the work of T7.3 is to find out what techniques, if any, are suited for supporting editors in their daily work of ontology maintenance (and engineers when creating new ontologies). The following dimensions should be taken into account:

1. precision
2. effort needed to apply the technique to the specific set of ontology and corpora (i.e. effort in input)
3. human effort needed to select the right suggestions.

Precision is a largely adopted measure in information retrieval and natural language processing. It is defined as the proportion of correct results out of the total results returned.

The effort needed to apply the technique includes any kind of pre-processing of the text corpus (or corpora) and/or ontology the system deals with. We call this “effort in input”.

The human effort to select the right suggestion is a more complex aspect to consider, as it depends on:

1. the number of appropriate suggestions (precision) provided by the system,
2. the quality of the ranking (good results should be placed at the top of the list),

3. the graphical interface used to display results to the user,
4. the interaction model implemented in the system to support the cycle of (i) inspecting candidates, (ii) selecting the good ones, (iii) place them in the appropriate position in the ontology.

In the work presented in this deliverable we mainly pay attention to item 1 and 2, while items 3 and 4 are used for the discussion taking place in Chapter 5.

1.2.6 Ideas for possible experiments

The following experiments are relevant applications to data dealt with in WP7.

1.2.7 Mapping AGROVOC/FAOTERM

Both resources are available already in the TBX common format. A mapping between FAOTERM subjects and AGROVOC has already been done. All FAOTERM entries belong to a subject which is an AGROVOC keyword (this may help realize the mapping itself). The current AGROVOC and FAOTERM web sites provide functionalities to make a search on both systems in order to retrieve relative information. Benefits are multiple: put together all information about a term. Get the corresponding definition of terms from FAOTERM and get more translations or relationships from AGROVOC. Make searches into databases indexed with AGROVOC to get relative documents while using FAOTERM, etc.

1.2.8 Mapping AGROVOC/CAT

The Chinese Agricultural Thesaurus (CAT) and AGROVOC have been manually mapped, according two these two types: ExactMatch and BroadMatch. The mapping has been done following specific guidelines and methodology. Both the two thesauri and the mapping are available in OWL.

1.2.9 Mapping AGROVOC/NAL

For this experiment AGROVOC and the National Agriculture Thesaurus (NALT) have been converted to SKOS. The two format in SKOS use a classifications scheme to organize the concepts in categories.

1.2.10 Mapping AGROVOC and ASFA

ASFA and AGROVOC have some overlapping, as AGROVOC includes a fragments about fishery-related topics, but they are also different in their structure (AGROVOC has a deeper hierarchy), and with respect to the multilinguality issue (AGROVOC is multilingual, ASFA is only in English). The alignment of ASFA and (the relevant fragment of) AGROVOC can result in the enrichment of ASFA with multilingual information.

1.2.11 Refining AGROVOC relationships

AGROVOC thesaurus has some basic relationships between terms (RT, USE/UF, etc.). Some relationships may be refined just using available document on those topics. The process can also be made (semi) automatically by adopting Soergel's "rules-as-you-go" approach (BIBREF). [1]

1.2.12 Learning from factsheets about fish stocks

Information about fish stocks (biological species living in a water area, e.g., "Bigeye Tuna, Pacific Ocean") is not explicitly stated in the reference tables for fisheries (BIBREF), but it can be extracted from reports and factsheets. The resulting extracted information can be used to enrich the existing ontologies (biological species and water bodies) with mappings across ontologies, or to create and populate an ontology about fish stocks.

1.2.13 Learning from factsheets about relations among fishing gears, water areas and biological species

In RTMS, AGROVOC and ASFA there are entries about fishing gears, water areas and biological species, but very little or no information about their relation: specific fishing gears are used to fish specific biological species in certain areas. This type of information, though, is contained in the factsheets, publications and the like, in an unstructured manner. The resulting extracted information can be used to populate existing ontologies, enrich AGROVOC or ASFA, or create links across ontologies.

1.2.14 Learning alternative mono- and multilingual names for fish

In RTMS, fish species are recorded together with their scientific name, and common names in English, French and Spanish. However, scientific names are not unique, and also in each language there are usually several alternative names for the same species. This type of information is contained in the FAO fact sheets, but also in the ASFA publications, and in other external resources: they could all be used to enrich ASFA/AGROVOC.

1.2.15 Learning from selected Fisheries related websites or full-text digital repository

The FAO document repository contains around ~5000 full-text documents related to the fisheries domain. These could be used to extract concepts (using NLP techniques) to populate fisheries ontologies. Similarly, pre-selected fisheries related websites could be also used to extract concepts.

[1] Some info on specific relationships extraction here: http://www.few.vu.nl/~wrvhage/pdf/iswc2006part-whole.pdf

2 State of the art in Ontology Learning

Ontologies are formal, explicit specifications of shared conceptualizations, representing concepts and their relations that are relevant for a given domain of discourse (Gruber, 1994). Currently, ontologies are mostly constructed by hand, which proves to be very ineffective and may cause a major barrier to their large-scale use in knowledge markup for the Semantic Web. Creating ambitious Semantic Web applications based on ontological knowledge implies the development of new, highly adaptive and distributed ways of handling and using knowledge that enable semi-automatic construction and refinement of ontologies.

Ontology Learning is an interdisciplinary area of interest, involving different scientific communities, such as Natural Language Processing, Knowledge Representation and Semantic Web. Therefore, defining a complete overview of the state of the art is not trivial. In addition, the methodologies elaborated since now are not very advanced, both from a theoretic and from a technological point of view. On the other hand, nowadays such technologies are becoming more and more interesting, due to the huge amount of knowledge required for the semantic web and to the increasing availability of non-structured information. Automation of ontology construction can be implemented by a combined use of linguistic analysis and machine learning approaches for text mining, which provides facilities for ontology construction and refinement (Maedche, 2003).In general, state of the art approaches do not reach the very high accuracy required to acquire formalized knowledge in a fully automated way. Therefore, only a semi automatic acquisition process can be reasonably attempted, where domain experts are asked to validate the domain knowledge proposed by the system.

In this section we will try to contextualize the experiments proposed in this document and to motivate the selection of the set of techniques described here. Ontology learning does not constitute a homogeneous area, since no formal theories and unified methodologies have been proposed yet. In contrast, it is a collection of techniques, very often mutated from previous research in Knowledge Acquisition and Information Extraction. A very high level partitioning of such techniques can be done by distinguishing all those methods relying on the analysis of textual material to those that requires already existing ontologies or partially formalized knowledge bases, such as dictionaries, thesauri, and so on. We will refer to the first group with the term “Ontology Learning from Text”, while the second group will be denoted “Knowledge Based Ontology Learning”. The following subsections will describe each group individually.

2.1 Ontology Learning from Texts

As human language is a primary mode of knowledge transfer, ontology development could be based more directly on the linguistic analysis of relevant documents. Most of the information contained in the web is expressed by natural language text. Human language is the primary vehicle for human communication, and can be used to refer to almost any conceivable fact and entity in any domain. Therefore, at least in principle, ontology learning from text should be possible, since the information contained in domain specific texts should express somehow the conceptualization we are looking for during the ontology engineering process. In fact, it is a common practice of ontology engineers to refer to textual documents when building the ontology.

On the other hand, many problems arise when trying to acquire knowledge from texts. Below we highlight the more relevant ones:

Ambiguity: words are very often ambiguous, while concepts in the ontology are supposed to be associated to a precise meaning

Variability: the same concept/entity could be referred to by different terms

Metaphora: Language is very often used in a metaphorical sense

Background Knowledge: Texts explicitly report only a minority of the actually intended knowledge, since their comprehension presupposes the availability of a shared conceptualization among sender and receiver.

Ambiguity and variability are very well known phenomena in lexical semantics. In general, ambiguity is sensibly reduced in domain specific texts, at least as far as domain specific terminology is concerned. In general, ambiguity affects precision, since irrelevant facts could be discovered regarding some concept/entity of interest in the domain. Variability is also a source of noise, since relevant facts involving some concept/entity of interest could not be recognized. On the other hand, the problem of variability is less relevant, since it mainly affects recall, which is not so relevant for the ontology learning purposes, where we are mainly interested in precision. A recent trend in computational linguistics is trying to deal with variability in order to solve the textual entailment problem.

The third problem is probably one of the hardest in the whole AI area since it has been largely studied achieving very partial and confusing results. The presence of metaphorical senses of terms in text could lead to the acquisition of false or irrelevant facts, since in general the system is not able to distinguish between metaphorical and standard senses of words, mainly affecting precision. In practice, most of the domain specific document collections in which we are generally interested in for the purposes of ontology learning are not that full of metaphors and creative usages of language, limiting the impact of metaphoric expressions in the acquisition process.

Problem 4) is probably the most relevant from the ontology learning point of view. Therefore it requires an ad-hoc discussion. In fact, the implicit knowledge presupposed by any writer when producing any text is exactly what we are looking for during the ontology engineering process. The background knowledge required for a speaker to understand any text is just rarely explicitly reported for economy reasons. Texts are just the tip of the iceberg of the huge amount of information they actually express. Acquiring background knowledge from texts is then a very hard problem, since it should deal primarily with paradigmatic (e.g. taxonomic) relations among words, which by definition are established “in absentia”, and therefore not easily captured with a shallow linguistic analysis (e.g. lexico-syntactic pattern based approaches). For this reason, deeper semantic processes should be preferable, for example the exploitation of distributional similarity measures for taxonomy induction, relation extraction and similarity.

A typical approach in ontology learning from text first involves the extraction of (more or less complex) terms from a domain-specific corpus. Extracted terms are statistically processed to determine their relevance for the domain corpus and clustered into groups with the purpose of identifying a taxonomy of potential classes. Additionally, relations can be identified, mostly by computing a statistical measure of connectedness between identified clusters. In the following subsections we will report the state of the art in each subtask of ontology learning in which we are interested for the purposes of this deliverable.

2.1.1 Preprocessing tools: GATE

The named entity and FAO species class recognition in T7.3 has been performed with GATE. GATE (http://www.gate.ac.uk/) is a world-leading system, which incorporates several text processing tools, such as tokenization, lemmatization, syntactic analysis and named entity

recognition. GATE is an architecture, a framework and a development environment for LE (Language Engineering) applications. As an architecture, it defines the organisation of an language engineering (LE) system and the assignment of responsibilities to different components. As a framework, it provides reusable implementations for LE components and a set of prefabricated software building blocks that language engineers can use, extend and customise for their specific needs. As a development environment, it helps its users minimise the time they spend building new LE systems or modifying existing ones, by aiding overall development and providing a debugging mechanism for new modules. Because GATE has a component-based model, this allows for easy coupling and decoupling of the processors, thereby facilitating comparison of alternative configurations of the system or different implementations of the same module (e.g., different parsers). The availability of tools for easy visualisation of data at each point during the development process aids immediate interpretation of the results.

Applications developed within GATE can be deployed outside its Graphical User Interface (GUI), using programmatic access via the GATE Application Programming Interface (API). In addition, the reusable modules, the document and annotation model, and the visualisation components can all be used independently of the development environment. GATE is engineered to a high standard and supports efficient and robust text processing.

The GATE architecture distinguishes between data, algorithms, and their visualisation. Following the terminology established in version 1, GATE components are one of three types:

1. LanguageResources (LRs) represent entities such as lexicons, corpora or ontologies;
2. ProcessingResources (PRs) represent entities that are primarily algorithmic, such as parsers, generators or ngram modellers;
3. VisualResources (VRs) represent visualisation and editing components that participate in GUIs.

These resources can be local to the user's machine or remote (available via HTTP), and all can be extended by users without modification to GATE itself.

For Named Entity Recognition (see section 2.1.3), Gate typically uses three types of processing resources: a gazetteer, a part of speech tagger and a rule grammar module. The gazetteer consists of lists such as cities, organisations, days of the week, scientific fish names etc. It not only consists of entities, but also of names of useful indicators, such as typical company designators (e.g. 'Ltd.'), titles, etc. The gazetteer lists are compiled into finite state machines, which can match text tokens.

The rule grammar component allows the encoding of rules that operate on the output of both the gazetteer and the pos tagger in order to annotate text spans with the relevant named entity types.

Chapter 4 describes into greater detail which tasks have been performed using Gate.

2.1.2 Terminology Induction

Terminology extraction, term extraction, or glossary extraction, is a subtask of information extraction. The goal of terminology extraction is to automatically extract relevant terms from a given corpus. If the corpus is a domain specific corpus, terminology extraction tools provide domain terms, which typically refer either to concepts or entities in a domain specific ontology.

Typically, approaches to automatic term extraction (Pazienza et al, 2005) make use of linguistic processors (part of speech tagging, phrase chunking) to extract terminological candidates, i.e. syntactically plausible terminological noun phrases, NPs (e.g. compounds "credit card", adjective-NPs "local tourist information office", and prepositional-NPs "board of directors" - in English, the first two constructs are the most frequent). Terminological entries are then filtered from the candidate list using statistical and machine learning methods. Once filtered, because of their low ambiguity and high specificity, these terms are particularly useful for conceptualizing a knowledge

domain or for supporting the creation of a domain ontology. Furthermore, terminology extraction is a very useful starting point for semantic similarity, knowledge management, human translation and machine translation, etc.

There exists a large number of tools for extracting terminology from corpora, some of them available as Web Services. See for example Term Extractor at http://lcl2.uniroma1.it/termextractor/ (Navigli and Velardi, 2004).

2.1.3 Named Entity Recognition

Named Entity Recognition (NER) is the task of identifying and categorizing entity names (such as persons, organizations, and location names), temporal expressions (dates and times), and certain types of numerical expressions (monetary values and percentages) in a written text. NER is a key part of information extraction but high performance systems also facilitate the annotation of corpora.

Systems for NER can be built based on deductive, knowledge-based methods such as list lookup and handcrafted rules, which usually rely on the combination of a wide range of knowledge sources (for example, lexical, syntactic, and semantic features of the input text as well as world knowledge and discourse level information) and higher level techniques (e.g. co-reference resolution). There are drawbacks related to the pure list lookup approach (Mikheev et al, 99), which mainly depend on the required dimensions of reliable gazetteers, on the difficulty of maintenance of this kind of resource, and on the possibility of overlaps among the lists. Moreover, their availability for languages other than English is rather limited.

Alternatively, NER techniques can be built on inductive techniques such as machine learning algorithms, or on hybrid methods (combinations of both) (Mikheev et al., 99).

Both deductive and inductive strategies utilize the so-called internal evidence, taken from within the NE, and the external evidence provided by the context in which a name appears (also called trigger words) i.e. predicates and constructions providing sufficient contextual information to determine the class of candidate proper nouns in their proximity. These trigger words can be defined manually, collected from the context of previously detected named entities in a text, or selected from existing lexical resources such as WordNet (Magnini et al, 2002). Given a set of labelled examples, external evidence, i.e. contexts and trigger words and internal evidence in the form of morphological or surface features such as capitalization can be learnt (McDonald, 96).

The inductive approach relies on the selection of features, and the learning of criteria for probable candidates on the basis of either a supervised setting, where the features are obtained from a gold standard, or an unsupervised setting, where the selection criteria are learned directly from text (quite often in combination with supervised methods, for instance linguistic pre-processing such as shallow syntactic parsing). Learners trained in these ways can be probabilistic (hidden Markov models, maximum entropy) (Cohn et al., 95), decision trees, support vector machines or memory-based (Daelemans and Van den Bosh, 2005).

In the supervised learning framework, a corpus of (typically) a few hundred documents is annotated by hand to identify the entities of interest. Features of local context are then used to train a system to distinguish instances from non-instances in novel texts. Such features may include literal word tests, patterns of orthography, parts of speech, semantic categories, or membership in special-purpose gazetteers.

Benchmarking and evaluations have been performed in the Message Understanding Conferences (MUC) organized by DARPA, International Conference on Language Resources and Evaluation (LREC), Computational Natural Language Learning (CoNLL) workshops, Automatic Content Extraction (ACE) organized by NIST, the Multilingual Entity Task Conference (MET), and the Information Retrieval and Extraction Exercise (IREX).

MUC (MUC-6, MUC-7) evaluations show that systems are able to score precision and recall values higher than 90% for English within a restricted domain.

Although highly successful in its application, there are still remaining issues for NER:

i) techniques for a cheap adaptation to new domains and new categories

ii) the development of effective systems for other languages, especially for languages where the characteristics of NER strongly differ from NER for English.

2.1.4 Relation Extraction

Relation extraction is a fundamental step in many natural language processing applications such as learning ontologies from texts (Buitelaar et al., 2005) and Question Answering (Pasca and Harabagiu, 2001).

The state of the art technology for relation extraction primarily relies on pattern-based approaches (Snow et al., 2006). These techniques, proposed by Hearst (1992), are based on the recognition of the typical patterns that express a particular relation in text (e.g. ***X such as Y*** usually expresses an **is-a** relation). In the literature several pattern based approaches have been proposed, some of them based on supervised techniques and therefore requiring manually annotated texts for each particular relation, others based on bootstrapping from a small set of seed patterns (Pantel and Pennacchiotti, 2006). The former approach requires a huge amount of work for annotation, even if it guarantees the best results as far as accuracy is concerned, while the second approach requires a minimal intervention, even if the achieved results are very often weak since the bootstrapping process could lead to some divergence. As a consequence, industrial scenarios very often adopt simple pattern based approaches, since the manual development of ad-hoc rules is not very expensive and the recognition algorithm can be easily implemented and maintained.

Yet, recent work (Gliozzo et. al, 2007) demonstrated that text-based algorithms for relation extraction, in particular pattern-based algorithms, still suffer from a number of limitations due to complexity of natural language, some of which we describe below:

- **Irrelevant relations.** These are valid relations that are not of interest in the domain at hand. For example, in a political domain, *"Condoleezza Rice is a football fan"* is not as relevant as ``*Condoleezza Rice is the Secretary of State of the United States*''. Irrelevant relations are ubiquitous, and affect ontology reliability, if used to populate it, as the relation drives the wrong type of ontological knowledge.
- **Erroneous or false relations.** These are particularly harmful, since they directly affect algorithm precision. A pattern-based relation extraction algorithm is particularly likely to extract erroneous relations if it uses *generic patterns*, which are defined (Pantel-and pennacchiotti,2006) as broad coverage, noisy patterns with high recall and low precision (e.g. "**X of Y**" for **part-of** relation). Harvesting algorithms either ignore generic patterns (Hearst, 1992) (affecting system recall), use manually supervised filtering approaches (Girju et al., 2006), or use completely unsupervised Web-filtering methods (Pantel and Pennacchiotti, 2006). Yet, these methods still do not sufficiently mitigate the problem of erroneous relations.
- **Background knowledge**. Another aspect that makes relation harvesting difficult is related to the nature of semantic relations: relations among entities are mostly paradigmatic (Saussure,22), and are usually established *in absentia* (i.e., they are not made explicit in text). According to Eco's position (Eco, 1979), the background knowledge (e.g. "*persons are humans*") is often assumed by the writer, and thus it is not explicitly mentioned in text. In some cases, such widely-known relations can be captured by distributional similarity techniques but not by pattern-based approaches.

- **Metaphorical language**. Even when paradigmatic relations are explicitly expressed in texts, it can be very difficult to distinguish between facts and metaphoric usage (e.g. the expression "*My mind is a pearl*" occurs 17 times on the Web, but it is clear that *mind* is not a *pearl*, at least from an ontological perspective).

Pragmatic issues (*background knowledge* and *metaphorical language*) and ontological issues (*irrelevant relation*) cannot be solved at the syntactic level, while they can be taken into account by adopting lexical distribution technique modelling semantic coherence through *semantic domains*. In particular, Gliozzo et. al (2007) demonstrated what they called "the domain restriction hypothesis", claiming that semantic relations can be established mainly among terms in the same Semantic Domain, while concepts belonging to different fields are mostly unrelated, showing that imposing domain restrictions to the candidate pairs of related entities effectively increases the precision of pattern based approaches.

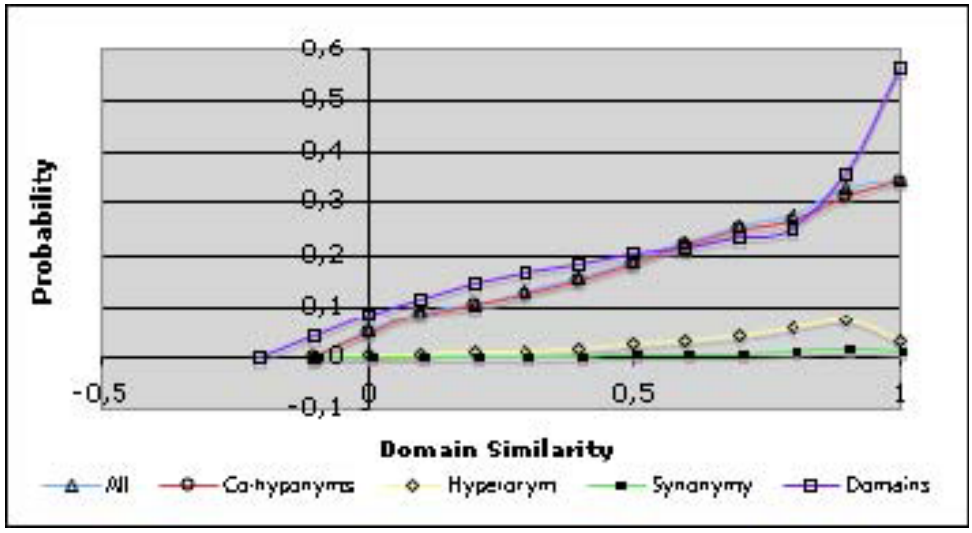

Figure 3. The domain restriction hypothesis

This hypothesis is supported by figure above (extracted from Gliozzo et. al (2007)), which shows the proportionality between the probability for two words to be related in WordNet with their domain similarity, measured in the LSA space induced from the British National Corpus. For each couple of words, the authors estimated the domain similarity, and collected word pairs in sets characterized by different ranges of similarity (e.g. all the pairs between 0.8 and 0.9). Finally, they estimated the probability of each couple of words in different ranges to be linked by a semantic relation in WordNet, such as *synonymy*, *hyperonymy*, *co-hyponymy* and *domain* in WordNet Domains (Magnini and Cavaglià, 2000). Results show a monotonic correspondence between these two quantities. In particular, the probability for two words to be related tends to 0 when their similarity is negative (i.e., they are not domain related), supporting the domain restriction hypothesis.

2.1.5 Ontology Learning Environments: Text2Onto

Text2Onto (Cimiano and Völker, 2005) is an ontology learning framework which has been developed to support the acquisition of ontologies from textual documents. Like its predecessor, TextToOnto (Mädche and Staab, 2004), it provides an extensible set of methods for learning atomic classes, class subsumption and instantiation as well as object properties and disjointness axioms. All of the algorithms being part of the Text2Onto framework largely rely on a combination

of machine learning and natural language processing techniques in order to extract ontology entities and relationships from open-domain unstructured text. Since the necessary linguistic analysis is done by means of GATE (Cunningham et al., 2002) it is very flexible with respect to the set of linguistic components used, i.e. the underlying GATE application can be freely configured by replacing existing components or adding new ones such as a deep parser if required. Another benefit of using GATE is the seamless integration of JAPE which provides finite state transduction over annotations based on regular expressions.

Linguistic preprocessing in Text2Onto starts by tokenization and sentence splitting. The resulting annotation set serves as an input for a part-of-speech (POS) tagger, which in the following assigns appropriate syntactic categories to all tokens. Finally, lemmatizing or stemming (depending on the availability of the regarding processing components for the current language) is done by a morphological analyzer or a stemmer, respectively. In order to improve the quality of the linguistic analysis particularly for Spanish text, some of the standard GATE components have been complemented by external resources. The TreeTagger is a POS tagger and lemmatizer developed by the University of Stuttgart which can be adapted to a multitude of languages by means of language-specific parameter files.

After the basic linguistic preprocessing is done, a JAPE transducer is run over the annotated corpus in order to match a set of particular patterns required by the ontology learning algorithms. Whereas the left hand side of each JAPE pattern defines a regular expression over existing annotations, the right hand side describes the new annotations to be created. Text2Onto makes use of JAPE patterns for both shallow parsing and the identification of modelling primitives, e.g. concepts, instances and different types of relations (Hearst, 1992).

In NeOn we are currently developing a graphical front-end for Text2Onto that will be made available as a plugin for the NeOn toolkit (see Figure 4). The plugin, which is going to be part of our prototype for learning networked ontologies (NeOn, D3.8.1), will enable the integration of Text2Onto into a process of semi-automatic ontology engineering.

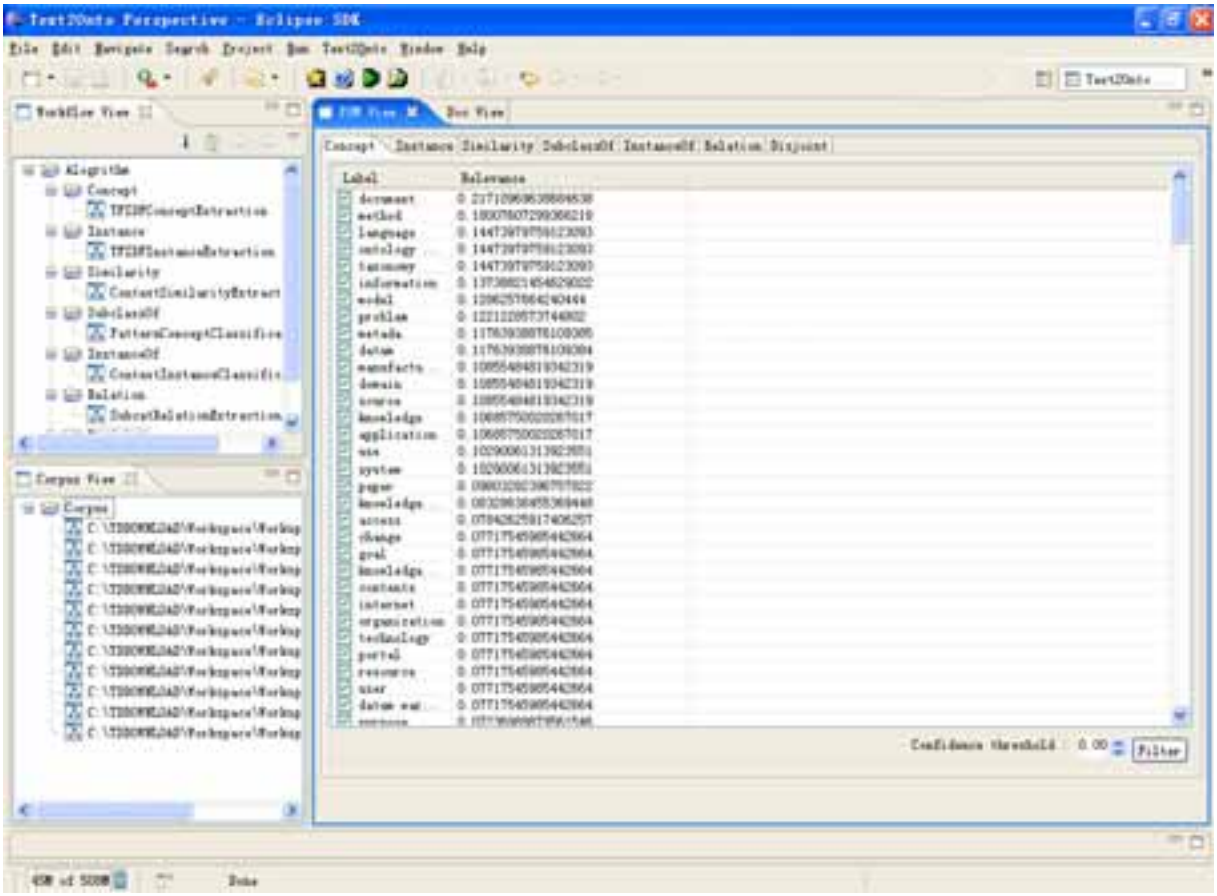

Figure 4. Text2Onto Plugin for the NeOn Toolkit

2.2 Knowledge Based Ontology Learning

Matching, mapping or aligning ontologies is a process where entities – mainly classes – of two different ontologies are linked together using ontological relations. In that sense, automatic ontology matching can be likened to the idea of ontology learning, as it focuses on discovering additional ontological knowledge expressed by ontological relations. In this section, we briefly overview the state-of-the art in ontology matching, focusing in particular on techniques that rely on the use of external background knowledge. In further sections, we experiment on the use of a technique relying on automatically selected ontologies to be used as background knowledge for matching (Sabou et al, 2005).

2.2.1 Ontology Matching

The issue of finding correspondences between heterogeneous conceptual structures is inherent to all systems that combine multiple information sources. The database community has identified schema matching as a core task in many application domains, such as integrating different databases (i.e., establishing mappings between their schemas), data warehousing and E-commerce (matching between different message schema) (Rahm and P. Bernstein 2001). Matching also plays a major role in approaches that rely on ontologies to solve the semantic heterogeneity problem between information systems (Kalfoglou and Schorlemmer, 2003) (Noy, 2004) (Wache et al., 2001).

The approaches developed both for database schemas and ontologies follow two major paradigms depending on the types of information they use to derive mappings (Kalfoglou and M. Schorlemmer, 2003) (Rahm and P. Bernstein, 2001) (Shvaiko and Euzenat, 2001). Internal approaches typically explore information provided by the matched ontologies such as their labels, structure or instances (Shvaiko and Euzenat, 2001). Indeed, all the ontology matching tools evaluated within the Ontology Alignment Evaluation Initiative (OAEI'06) primarily exploit label and structure similarity to derive correspondences associated to varying confidence values (Euzenat et al., 2006). A limitation of such approaches is that they depend on the richness and the similarity of the internal information of the matched ontologies. For example, Aleksovski et al. (2006) used two state of the art matchers, FOAM (Ehrig and Y. Sure, 2005) and Falcon-AO (Jian et al, 2005), to match weakly structured medical vocabularies with a low overlap in their labels and obtained precision values of only 30% and 33%.

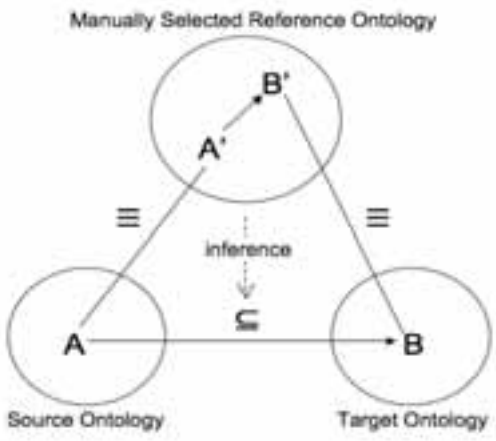

Figure 5 Using background knowledge for ontology matching

External (or background knowledge based) techniques aim to address this limitation by exploring an external resource to bridge the semantic gap between the matched ontologies. Indeed, continuing the example above, Aleksovski et al. (2006) obtained a precision value of 76% on the same dataset in the medical domain by exploring the DICE ontology as background knowledge (Aleksovski et al, 2006). As depicted in Figure 5, matchers from this category exploit an external resource by replacing the original matching problem (between concepts A and B) with two individual matching and an inference step: the two concepts are first matched to so called anchor terms (A', B') in the background source, and then mappings are deduced from the semantic relations of these anchors.

2.2.2 Ontology Matching based on Background Knowledge

We distinguish two categories of matchers relying on external background knowledge, depending on the type of the explored external resource, i.e., an ontology (Aleksovski et al, 2006), (Bouquet, 2005), (Collet et al, 1991), (Stuckenschmidt et al, 2004) or online textual sources (Van Hage et al, 2005).

Several ontology based matchers rely on a large-scale generic resource such as Cyc or WordNet. The Carnot system (Collet et al, 1991), (Huhns et al,1993) explores the Cyc knowledge base as a global context for achieving a semantic level integration of various information models (e.g., database schemas, knowledge bases). CTxMatch (Bouquet, 2005) (and its follow-up, SMatch (Giunchiglia et al, 2005)) translates ontology labels into logical formulae between their constituents,

and maps them to the corresponding WordNet senses. A SAT solver is then used to derive mappings between the concepts. This approach has been extended to handle the problem of missing background knowledge (Giunchiglia et al, 2006): if the simple techniques used to explore WordNet fail, then a second set of more complex and computationally expensive heuristics are applied to gain more knowledge.

While readily available, generic resources might fail to provide the appropriate coverage when matching is performed in a specific domain, such as medicine. In these cases, several matching approaches have opted for the use of a domain ontology. The SIMS system (Arens et al, 1993). relies on a manually built ontology about transportation planning for integrating several databases in this domain. In (Aleksovski et al, 2006), the authors match two weakly structured vocabularies of medical terms by using the DICE ontology. Similarly, in (Stuckenschmidt et al, 2004) mappings between two medical ontologies (Galen and Tambis) are inferred from manually established mappings with a third medical ontology (UMLS), and by using the reasoning mechanisms permitted by the C-OWL language. Unfortunately, building (and even selecting) an appropriate domain ontology prior to matching is a considerable effort and represents a drawback of these techniques (Arens et al, 1993).

Van Hage et. al (2005) use the combination of two “linguistic ontology matching techniques” that exploit online texts to resolve mappings between two thesauri in the food domain. First, they rely on Google to determine subclass relations between pairs of concepts using the Hearst pattern based technique introduced by the PANKOW system (Cimiano et al, 2004). Then, they exploit the regularities of an online cooking dictionary to learn hypernym relations between concepts of the matched ontologies. The strength of this approach is that, in principle, it is domain independent and therefore it does not require manual background knowledge selection. In reality, however, its precision dramatically decreases when relying on a corpus of general texts (50%), as opposed to a domain specific one (75%).

3 Resources and Tools

This section describes the resources used for the experiments described in Chapter 4. These resources consist of three different types:

1. the resources provided by FAO in the form of ontologies and text corpora,

2. the results of the pre-processing of the text corpora by USFD in the form of segmentation and linguistic/terminological annotation of the texts;

3. the interfaces created by OU and USFD for the manual evaluation of the experimental results performed by FAO.

Each of these will be described in a separate section below.

3.1 Available Resources

The experiments investigate the feasibility of improving existing ontologies in the fisheries domain by using text mining tools (see Chapter 4 for detailed descriptions of these experiments).

To accomplish this goal we need reference ontologies, to be used either as a gold standard or as a core ontology to be further extended, and large collection of domain specific documents. For the WP7.3 experiments we concentrated on the Fisheries domain, by selecting domain specific corpora and ontologies to be used for the experiments. Below we describe the existent resources currently in place at FAO and the processing we did in order to make them interoperable and shared among partners

3.1.1 Corpora

The corpora have been collected by FAO and put at the disposal of the other partners. All of them are briefly described below, although the experiments have concentrated on corpora 2 and 3.

3.1.1.1 Fisheries Atlas CD

This is the collection of texts selected form the Fisheries Atlas CD provided by FAO. It can be downloaded from http://www.loa-cnr.it/Files/NEONWP73/test.zip

3.1.1.2 FIGIS fact sheets

This corpus is 330 MB, and contains around 2M words in 1383 XML fact sheets, which describe aspects of fish species in free and semi-structures text form. It has been compiled from the FAO data disc "FI resources". This disc contains a list of urls organized under the following headers:

- country_sector_factsheet_urls
- culture_species_factsheet_urls
- data_collection_factsheet_urls
- fisheries_orgs_factsheet_urls

- fisherires_topics_factsheet_urls
- fishery_area_factsheet_urls
- fishery_factsheet_urls
- fishing_equipment_factsheet_urls
- geartype_factsheet_urls
- legal_framework_factsheet_urls
- resource_factsheet_urls
- species_factsheet_urls
- standards_factsheet_urls
- vesseltype_factsheet_urls
- vms_programme_factsheet_urls

The corpus was compiled by crawling these urls.

3.1.1.3 FAO corporate document repository (FCDR)

This corpus is 183MB, and contains 4963 documents in the fisheries domain collected from the FAO website.

Original texts (.txt format) can be downloaded from http://www.loa-cnr.it/Files/NEONWP73/fao-corporate-document-repository-text.txt.tgz

3.1.1.4 ASFA abstracts

This corpus consists of all English abstracts and related metadata (367.696 records) from the ASFA system in textual format. It can be downloaded from http://ponta.ijs.si/pub/asfa/asfa-English.zip (240 MB).

3.1.2 Ontologies

This section will only describe the ontologies developed by FAO, which have directly or indirectly been used in the experiments.

3.1.2.1 Water Bodies ontology

This ontology has been generated from FIGIS data, stored in 3 tables (fic_catch_area, fic_catch_area_agg_grp, and md_refobject) with ODEMapster (see D7.2.2)

The ontology model is as follows:

class: FAO fishery water body consists of 4 levels of classes: AREA, SUBAREA, DIVISION, SUBDIVISION.

datatype property: each class has datatype properties of a) 3 official languages of FAO (English, Spanish, and French), b) geographical information (latitude and longuitude) and c) size and d) code for UN, UNDP, ISO2,ISO3,and FAO code.

object property:

- An object property called "hasSubArea" has a domain "AREA" and a range "SUBAREA".
- An object property called "hasDivision" has a domain "SUBAREA" and a range "DIVISION".
- An object property called "hasSubDivision" has a domain "DIVISION" and a range "SUBDIVISION".
- An object property called "isInArea" has a domain "SUBAREA" and a range "AREA", which is an inverse property of the property "hasSubArea".

- An object property called "isInSubArea" has a domain "DIVISION" and a range "SUBAREA", which is an inverse property of the property "hasDivision".
- An object property called "isInDivision" has a domain "SUBDIVISION" and a range "DIVISION", which is an inverse property of the property "hasSubDivision".

Number of generated instances: AREA (28) SUBAREA(93) DIVISION(10) SUBDIVISION(29)

The class labels from this ontology, which mainly consist of abbreviations and numbers (e.g. “E Indian O.”, “Europe inl”, “34.2”) are provided with mappings onto the FIGIS LargeMarineEcosystem class, which contains text labels, and has been used for the annotation of the texts. This FIGIS list contains 50 elements, of which 28 feature in the evaluation experiment:

- Agulhas Current
- Arabian Sea
- Baltic Sea
- Barents Sea
- Bay of Bengal
- Benguela Current
- Canary Current
- Caribbean Sea
- Celtic- Biscay Shelf
- East China Sea
- Eastern Bering Sea
- Gulf of Alaska
- Gulf of California
- Gulf of Mexico
- Humboldt Current
- Iberian Coastal
- Mediterranean Sea
- New Zealand Shelf
- North Sea
- Norwegian Shelf
- Red Sea
- Scotian Shelf
- Sea of Japan
- Somali Coastal Current
- South China Sea
- Southeast U . S . Continental Shelf
- Sulu- Celebes Sea
- Yellow Sea

3.1.2.2 Biological species

The ontology model is as follows:

Classes:

- Classification
    - group

    - order
    - family
    - species
- Class restrictions: universal
- group
    - forAll includesOrder order
    - forAll includesFamily family
    - forAll includesSpecies species
- order:
    - forAll includesFamily family
    - forAll includesSpecies species
- family
    - forAll includesSpecies species

Datatype properties:

- Domain = Classification:
- hasNameEN type:string xml-lang:en
- hasNameES type:string xml-lang:es
- hasNameFR type:string xml-lang:fr
- hasLongNameEN type:string xml-lang:en
- hasLongNameES type:string xml-lang:es
- hasLongNameFR type:string xml-lang:fr
- hasFullNameEN type:string xml-lang:en
- hasFullNameES type:string xml-lang:es
- hasFullNameFR type:string xml-lang:fr
- hasTaxCode
- hasMeta
- hasISSCAPcode

Object properties

- includesSpecies = Domain: classification, Range: species
- includesOrder = Domain: classification, Range: species
- includesFamily = Domain: classification, Range: species

Details on this ontology can be found in D7.2.2.

3.1.2.3 Agrovoc

AGROVOC is a multilingual structured thesaurus of all subject fields in agriculture, forestry, fisheries, food and related domains (e.g. environment). It consists of words or expressions (terms),

in different languages and organized in relationships (e.g. “broader”, “narrower”, and “related”), used to identify or search resources. Its main role is to standardize the indexing process in order to make searching simpler and more efficient, and to provide the user with the most relevant resources.

The AGROVOC Thesaurus was developed by FAO (Food and Agriculture Organization of the United Nations) and the Commission of the European Communities, in the early 1980s. It is updated by FAO roughly every three months, a and the user can see the specific changes on the AGROVOC website (http://www.fao.org/aims/ag_intro.htm). Currently the thesaurus contains around 38,000 terms.

AGROVOC is used all over the world, and available in the five official languages at FAO, which are English, French, Spanish, Chinese and Arabic. It is also available in Czech, Portuguese and Thai. Other languages such as German, Italian, Korean, Japanese, Hungarian, and Slovak, are currently either being translated or revised.

3.1.2.3 NALT

The United States National Agricultural Library (NAL) Agricultural thesaurus NALT (http://agclass.nal.usda.gov/agt/agt.shtml) contains more than 62,000 agricultural terms.

NALT is primarily used for indexing and for improving retrieval of agricultural information. Currently, NALT provides the indexing vocabulary for NAL's bibliographic database of citations to agricultural resources, AGRICOLA. The Food Safety Research Information Office (FSRIO) and Agricultural Network Information Center (AgNIC) also use the NALT as the indexing vocabulary for their information systems. In addition, the NALT is used as an aid for locating information on the ARS and AgNIC web sites.
The Thesaurus is organized into 17 subject categories, indicated by the "Subject Category" designation in the thesaurus. Use the subject categories and other topics to browse the Thesaurus in a specific discipline or subject area.
The subject scope of agriculture is broadly defined in the NALT, and includes terminology in the supporting biological, physical and social sciences. Biological nomenclature comprises a majority of the terms in the thesaurus and is located in the “Taxonomic Classification of Organisms” Subject Category. Political geography is mainly described at the country level.

3.1.3 GATE pre-processing and Named Entity Extraction

The University of Sheffield pre-processed three corpora: the FIGIS fact sheets, the FAO document repository and the ASFA abstracts. First, structural and linguistic pre-processing was performed by means of GATE (see Section 2.1.1):

- Tokenization
- Orthographic analysis (e.g. capitalization).
- Word length
- Sentence splitting
- Part of speech tagging
- Named Entity Recognition on the basis of gazetteer lookup and heuristic rules.

For the last task, gazetteer lookup was used for the annotation of text spans with ontological classes from the FAO ontologies. Some lemmatization was performed by the normalization of some nominal inflectional paradigms (e.g. plural –s and Latin masculine plural –i for singular –us).

On top of this, some heuristic rules were applied to discover terminology not yet present in the ontologies.

This resulted in the following Named Entity classes (NEs):

- General language use:
  - FishClass from WordNet (all text instances identical to any WordNet synset member from synsets that are hyponyms of the synset containing “fish” (any of various mostly cold-blooded aquatic vertebrates usually having scales and breathing through gills; “the shark is a large fish”; “in the living room there was a tank of “colourful fish").
  - FishGeneric (the literal occurrence of “Fish” or “fish” in text)
- FAO NEs:
  - FishClass from the species ontology
  - FishClass from Agrovoc
  - LargeMarineEcosystemClass (queried from the FIGIS database)
- Gate NEs:
  - Address
  - Date
  - FirstPerson
  - Person
  - JobTitle
  - Location
  - Money
  - Organization
  - Percent
  - Title

The annotated FIGIS Fact Sheet corpus can be downloaded from: http://www.loa-cnr.it/Files/NEONWP73/factsheets-xml-fi-ag-wn.zip .The annotated FAO Document Repository can be downloaded from: http://www.loa-cnr.it/Files/NEONWP73/fao-corporate-document-repository.zip.

The annotations can be visualized in the GATE GUI (downloadable from http://www.gate.ac.uk/),

Figure 6 below shows a screenshot from the GATE GUI. Annotation classes are in the right-hand pane. Text elements annotated with “Fishclass” have been selected to be shown in the middle pane. Additional attributes can be seen in the inset window, which pops up when the cursor is directed onto the text element. The most interesting attribute in this example is the name of the family to which “Genypterus Blacodes” belongs (“Cusk-eels brotulas nei”).

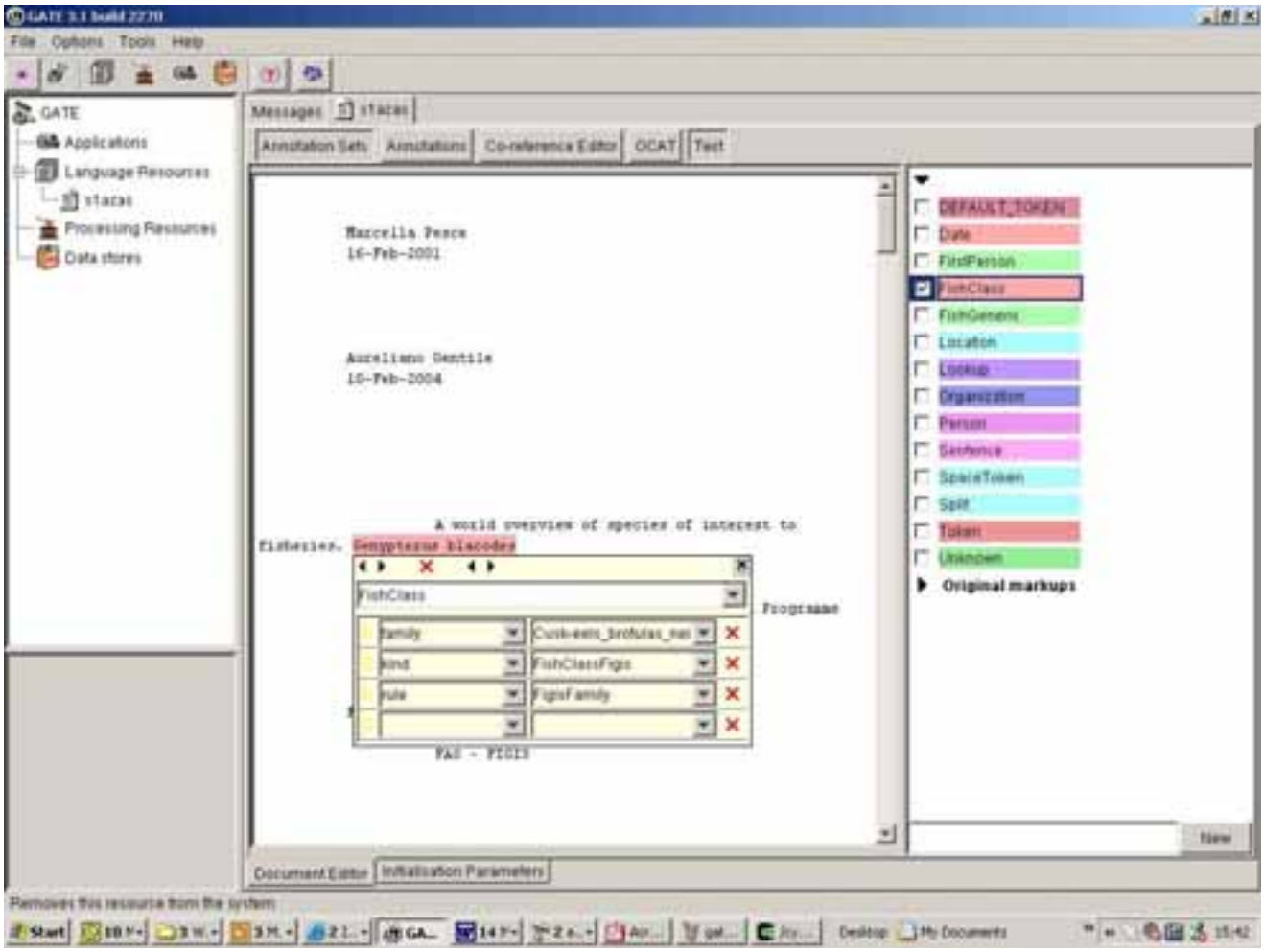

Figure 6. Example of text annotationin GATE

The evaluation of the results of the entity annotation process have not been taken into account in the experiments for the following reasons:

- The entities derived from the FAO ontological and terminological resources contain mostly unambiguous scientific vocabulary, and therefore the quality of the annotations is expected to be very high;
- The named entities extracted by the general purpose GATE NE system mostly fall outside the scope of the FAO ontologies, and can only be taken into account for ontology extension and mapping purposes in the future. Examples are Person and Organization.
- The domain-specific fishery term candidates that have been found by means of the application of pattern-based rules do fall within the scope, and deserve evaluation in the near future, as this can be seen as an extension of the successful pattern-based extraction performed by UKARL.

Until now, the following heuristics have been implemented:

- Candidate subclasses of existing elements from the species ontology:

  - Adjective-noun classes where nouns exists as labels in the species ontology, e.g. “Japanese flounder”
  - Heart patterns: “especially”, “such as”, “including”, “and other”, “or other”
  - Headword mapping, e.g. canddidate “Suberites sponges” as hyponym of existing concept label “sponges”

- Candidate synonyms: bracketed mentions of recognized fish names indicate synonymy, e.g. “Mummichogs (Fundulus Heteroclitus) were the most common single prey item…”
- Candidate Fishclass: list membership e.g. “Tuna, clamps and herring”

The first version of this application, which performs conceptual indexing of texts and suggests candidate terminology, is available as a Gate annotation web service (called “Sardine”), executable through the Gate NeOn toolkit plug-in.

In general, proper evaluation is required, because it has to take into account that the GATE-derived entities will not always be correct, because terminology of names of e.g. places and persons is much more ambiguous, and heuristic rules are not always successful.

In the case of WordNet, which covers ordinary names as well as some scientific terminology, errors caused by ambiguity of the terms occur. For instance, the synset {drum, drumfish} (small to medium-sized bottom-dwelling food and game fishes of shallow coastal and fresh waters that make a drumming noise) causes the incorrect annotation of “drum” in e.g. “revolving barrel or drum incubator”. In order to address this issue, further disambiguation strategies need to be integrated.

3.1.4 Evaluation Interfaces

According to the user requirements reported in Chapter 1, the validation and evaluation process has been performed by domain experts by adopting easy and intuitive interfaces, with the goal of minimizing the evaluation effort and to make the process more effective. Below we describe the evaluation interfaces we developed for the experiments reported in Chapter 4.

3.1.5 Relation Extraction

For evaluating the relation extraction experiments we build evaluation interfaces as Microsoft Access forms. The evaluators were asked to either tick boxes or fill in values to indicate their evaluation. Comment fields allowed any remarks from the side of the evaluators.

Figure 7 below illustrates the interface developed to evaluate the relations between water bodies (LargeMarineEcosystem) and fish species by adopting the LSA based relation extraction method. The evaluator has scored the relatedness between “Carribean Sea” and “Pacific Seabob” (with a latent semantic indexing score of 0.53) as “No”. The evaluation is assisted by the “CONTEXT” panes, which show keyword-in-context instances of the water body and the fish respectively.

Within each of these contexts, “Field2” contains the source file from which the context in Field3 has been extracted. Multiple occurrences within the same file are listed separately (see the “CONTEXT fish” pane).

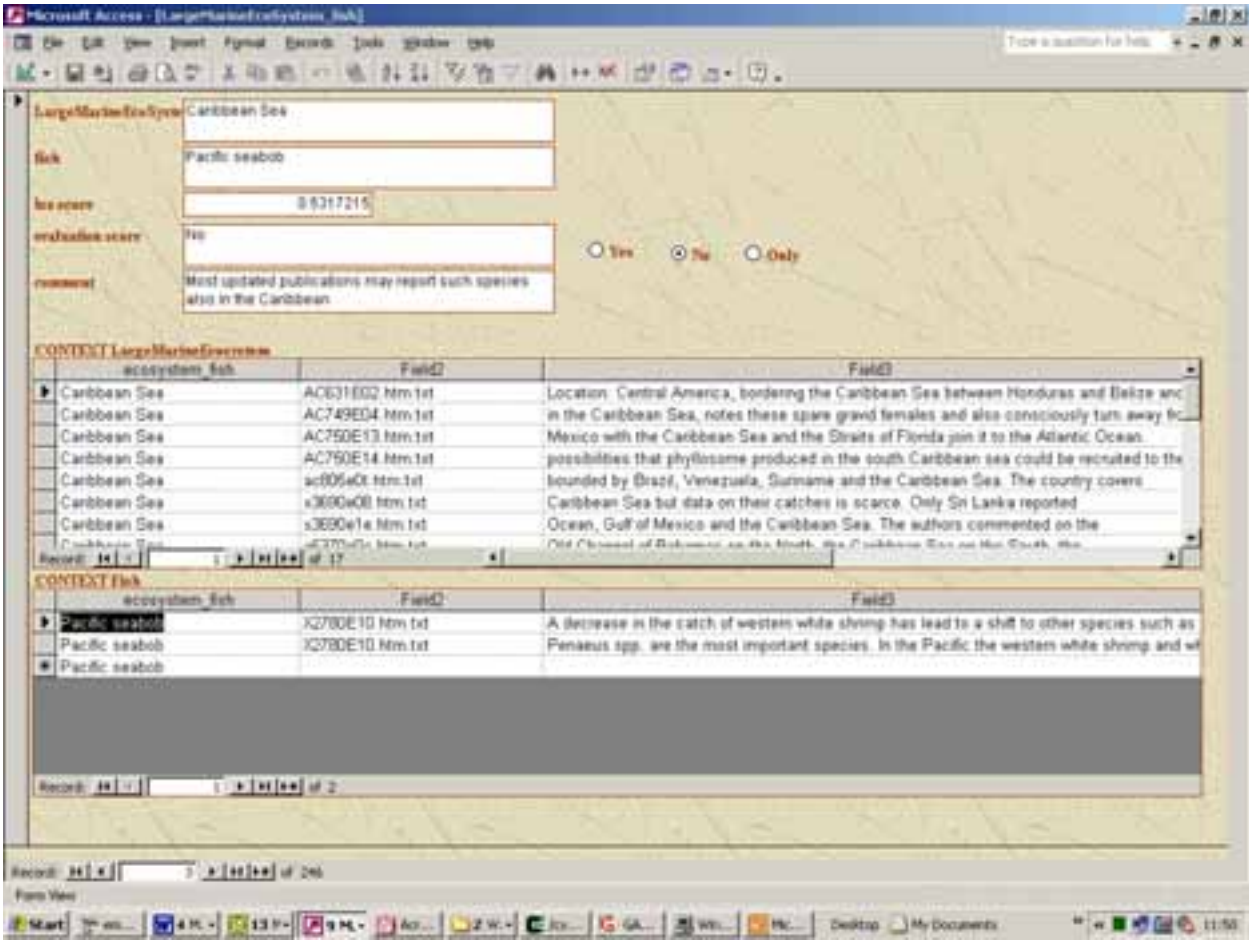

Figure 7: Evaluation interface adopted for LSA based relation extraction

Figure 8 illustrates the evaluation of the results of the FIGIS fact sheet relation extraction experiment. The candidate “lives in” relations between species and location have been manually evaluated with help from a context pane. Many of the species classes have synonyms associated with them in the fact sheets. The correctness of these synonyms has been scored per species on the bottom half of the form.

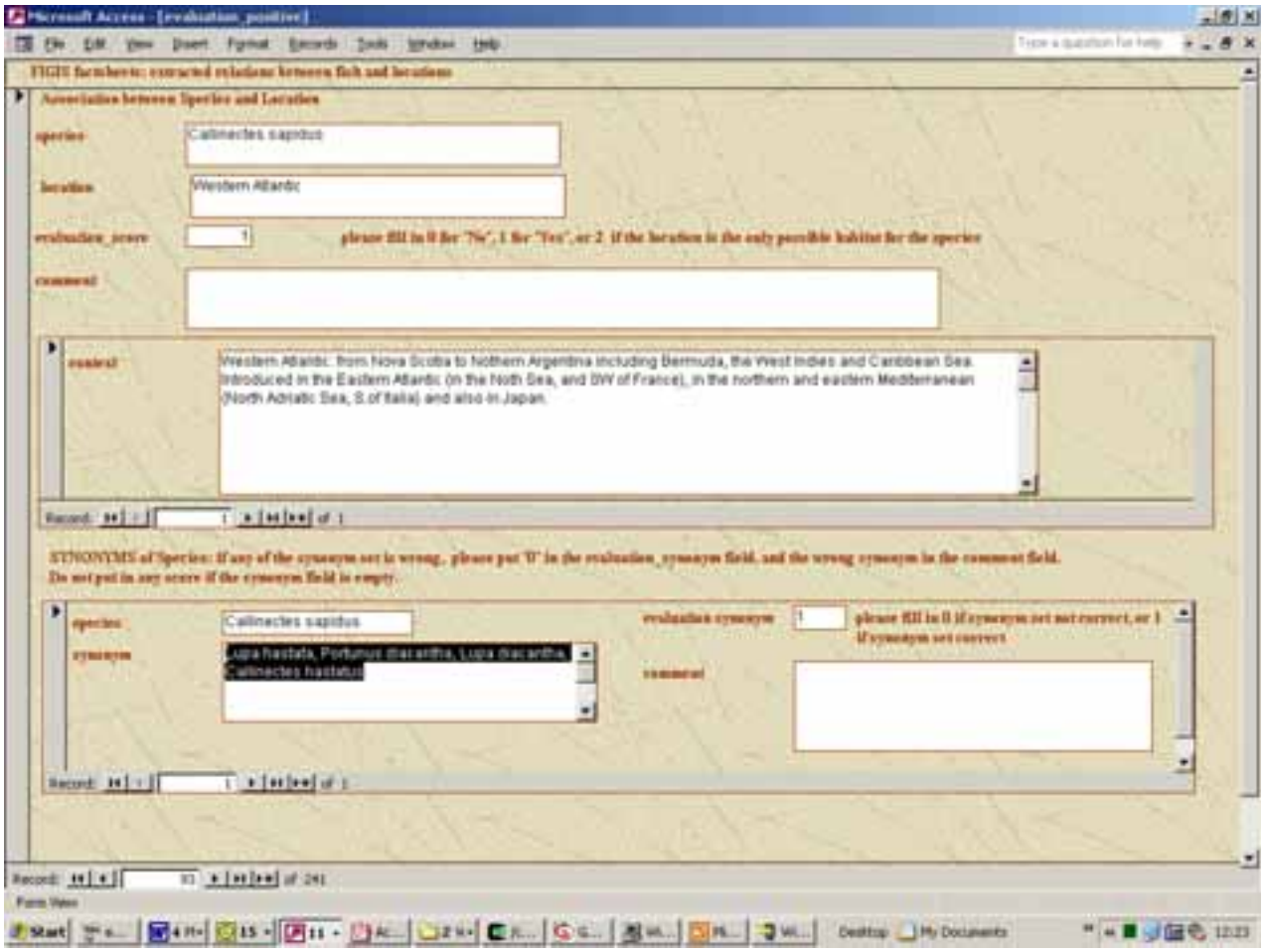

Figure 8: Evaluation interface adopted for the pattern based relation extraction experiments

3.1.6 Ontology Mapping

For the purpose of the evaluation of the obtained results, we have developed an evaluation interface and a guideline for evaluators. Evaluating the whole set of mappings generated by our approach (3,500 mappings) would be too expensive. For this reason, we divided these results into samples. We believe that analyzing the evaluation of a sufficient number of samples would give results general enough to extrapolate on the whole set.

The evaluation interface is a simple application that can be loaded with a set of mappings for which different evaluators can provide an evaluation. The evaluation for each mapping consists on the answer to the question “is this mapping correct?”, the possible answers being yes, no, or “I don’t know”, that can be completed with an optional comment (free text). Figure 9 shows what the main window of this interface looks like.

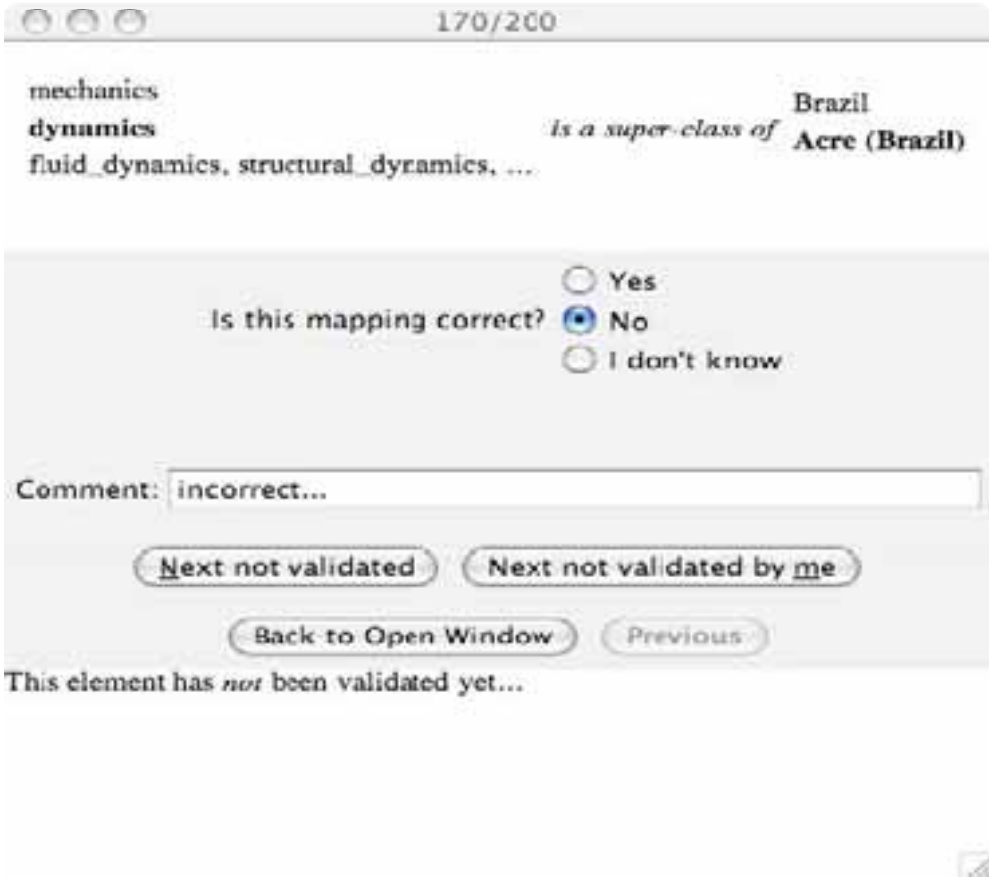

Figure 9. The evaluation interface for a super-class mapping

3.1.6.1 Interpretation of the mapping relation.

One of the common mistakes in mappings is a misinterpretation of the mapping relation linking two concepts. Therefore it is important, when evaluating mappings, to have a good understanding of these relations.

- *sub-class-of* : A concept, or a class, represents a set of entities having common properties. A concept C is a sub-class of another concept D if it represents a sub-set of the set represented by D. In other terms, we can say that any C is a D. For example, a sub-class mapping between Cat and Feline would be correct, since any Cat is a Feline. The inverse mapping (Feline sub-class-of Cat) would be incorrect since a feline is not necessarily a cat. The sub-class-of relation is often confused with other relations like part-of or related-to. For example, China sub-class-of Asia is incorrect: China is a part of Asia. The sentence “any China is an Asia” does not make any sense.
- *super-class-of*: This relation is the inverse of the sub-class-of relation: C super-class-of D is equivalent to D super-class-of C . For example, Feline is a super-class of C.
- *equivalent-to*: Two classes are equivalent if they represent exactly the same set of objects, so if they are at the same time sub-classes and super-classes of each other.
- d*isjoint-with*: Two concepts are disjoint if the sets of objects they represent do not have any overlap. In general, C is disjoint with D means that no object from C can be an object from D. For example, Cat and Dog are disjoint concepts, since an object being a cat cannot also be a dog.

3.1.6.2 A mapping has to be correct in the context

One of the common mistakes in ontology matching comes from the ambiguity of the words used to describe concepts: a word may have different meanings in different contexts. For example, you may find a mapping saying that RAM is a sub-class of Memory. This mapping is correct in the context of computer components, where RAM is a memory device, but is incorrect if we talk about ram as a male sheep. The validation interface gives an indication of the context in which a word has to be interpreted by providing the super-classes and sub-classes of the considered concepts in the source and target ontologies.

3.1.6.3 What if you don't know?

There will also be several mappings that might be impossible to evaluate because you do not have the proper domain knowledge about the concepts to be mapped (e.g., that leukemia is a neoplasm). In these cases you can use any external resource to get a better understanding about how the concepts relate (e.g., online dictionaries, Google - checking pages where these concepts occur together). Another good resource to check is WordNet (both the definitions of the concepts and the hypernymy information available), for those of you familiar with it. If you cannot find any info about these two concepts in a reasonable amount of time, then you can simply press the "I don't know" button.

4 Techniques and Evaluation

This section describes the techniques we adopted for our ontology learning experiments and reports the evaluation we performed. Since we experimented with very different techniques, we will report each of them individually in the following subsections. Further analysis of the achieved results will be reported in Section 5, where the most promising ones will be recommended for integration in the ontology lifecycle.

4.1 Terminology Extraction

The goal of this set of experiments is to evaluate the state of art terminology extraction technology when applied to the extraction of terms from Fisheries related websites or full-text digital repository. This terminology will be used to suggest new concepts/terms/instances to domain experts which will validate them in order to populate the ontology.

4.1.1 Use case

The domain expert is interested either in creating a new ontology or in populating an existing one with novel concepts and instances. To this aim he submits a corpus of domain specific texts to the system. As an output, the system provides a ranked list terms extracted from texts. The ranking is decided according to a reliability function.

4.1.2 Description of the techniques

For the terminology induction experiments we compared two different techniques: a more standard based on regular expressions and statistical measures of reliability (Navigli and Velardi, 2004) and a novel technique based on the exploitation of SuperSenseTagging (SST) (Ciaramita and Altun, 2006) (Picca et al, 2007).

Since the former techniques are more standard (see the state of the art section), there exist many freely available tools that implement them. For our experiments we adopted Term Extractor (http://lcl2.uniroma1.it/termextractor/) since it can be used as a WEB server, it is able to work with texts in multiple formats (pdf, txt, doc among the others) and it is provided by a nice user interface. The output of this system is a ranked list of terms, where the ranking is selected according to a reliability function depending two parameters: (i) the internal coherence between words composing the term measured by their mutual information and (ii) the degree of representativeness of the term for the domain specific corpus. As an example, below we list the top ranked terms returned by term extractor on the fishery domain.

Resource
Management
Forest
Fishery
united nation
Ffish

Fao
Species
case study
Policy
Data
Africa
Problem
developing country
Fishing
Percent
Agriculturae
Forestry
Method
Approach
Department
Member

As an alternative approach, we experimented with the exploitation of the SST technology for terminology extraction. Since this technique is less standard, we implemented an ad-hoc prototype system for the experiments, following the methodology described in Picca et al (2007), and adapting it to our use case. In particular we adopted SST as a preprocessing system to identify concepts and instances in texts. SST, broadly discussed in Ciaramita and Altun (2006), is the problem to identify terms in texts, assigning a "supersense" category (e.g. *person*, *act*) to their senses in context. In analogy to a standard terminology extraction system, SST identifies terms in texts. In addition to standard terminology extraction technology, SST recognizes the high level *ontological type* for each term, selected from a repository of about 20 high level supersenses in WordNet. Previous research conducted at CNR demonstrate that such supersenses cover most of the typically used high level categories in ontology design patterns, making the supersenses the right level of abstraction for ontology design. SST is based on a supervised Named Entity Recognition (NER) technology for entity boundaries detection. In contrast to standard NER, SST identifies both concepts and entities in texts, while the former is only able to identify entities, when adequately trained. In addition, SST does not require any labelled data for training or domain adaptation, because its categories are totally general and valid for any domain. As output the SST returns a tokenized text where each token is annotated with either a B-X , I-X or O-X tag, indicating the beginning , the internal part or the absence of a term in the sentence. As an output we obtained lists of categorized terms, lemmatized and ranked by their frequency. As an example, below we report the top ranked part of the list belonging to the type *location*.

Thailand
Indonesia
Country
Philippines
District
China
Trend
State
Border

Since the type distinction provided by SST is very coarse grained, the extracted terms should be further subcategorized into the finer grained concepts represented in the ontology. Anyhow, the coarse grained categorization can be perceived as a first filtering step, which can be performed in a totally automatic and domain independent way.

4.1.3 Evaluation

4.1.3.1 Experimental settings

For our experiments, we run both TermExtractor and SST on the FAO Corporate Document Repository, and we submitted the extracted term lists to domain experts. Term extractor identified around 6000 terms, out of which we selected the top ranked 300 to be validated by domain experts.

SST extracted different ranked lists of each ontological type. In total we extracted 17 lists of terms, described reported in the following table.

Table 1. Number of terms extracted from the FAO corporate document repository for each category

Category	Number of terms
Animal	4116
Artifact	17643
Body	1913
Cognition	6048
Communication	11706
Event	1505
Feeling	665
Food	2477
Group	52558
Location	17369
Person	57857
Plant	1994
Quantity	3543
Relation	831
State	3750
Substance	4956
Time	2946

Each list contains either terms denoting concepts or terms denoting instances. The ranking criterion based on frequency tends to put in the top ranked part of the list very abstract concepts, leaving at the bottom very specific ones and most of their instances. For example, instances of sharks in the list have the following frequency.

Shark	66
shark_fin	17
shark_fishery	16
shark_catch	15
mako_shark	5
shark_consumption	4
shark_export	4

dog_shark	4
shark_fins	3
shark_cartilage	3
shark_utilization	3
tail_shark	3
catsharks	2
bull_shark	2
shark_bycatches	1
elasmobranch_catshark	1
Survey_shark	1
fish_catshark	1
freeze_shark	1
indian_shark	1
shark_bile	1
shark_hide	1
shark_meatball	1
shark_importer	1
nurse_shark	1
liveroil_shark	1
tope_shark	1
cape_shark	1
Dogfish_catshark	1
deepwater_shark	1
spadenose_shark	1
catshark	1

As expected the concept of shark has the highest frequency, while its particular instances are more rare, and then placed in the bottom part of the list.

Out of those 17 ontological types, we submitted the domain experts only the following, since a preliminary investigation about revealed that they are the more relevant for the application domain:

- Food
- Plant
- Animal
- Location

We also evaluated the type *substance*, which is clearly not relevant to the domain, as a proof of concept, expecting that less relevant terms will be returned.

Being the full lists of retrieved terms for each type too long to be fully validated, we submitted to the domain expert random samplings of those lists, each composed of 30 terms, whenever available. As a sampling criterion, we selected the following frequency ranges:

- Above 1000 occurrences
- Between 101 and 1000 occurrences
- Between 11 and 100 occurrences
- Between 2 and 10 occurrences
- Only 1 occurrence (singletons)

Overall, we collected around 120 terms per ontological type (in many cases, terms above 1000 occurrences where less than 30 due to the small corpus dimension), and we submitted them to domain experts for validation.

4.1.4 Annotation Guidelines

Both from the population and from the conceptualization point of view, the principal requirements of a terminology extraction system are essentially accuracy and recall. In fact, if the percentage of out-of-domain and syntactically non-well-formed terms is too high, the effort required to identify errors from the list would become higher than the benefits. In practice, recall is very difficult to estimate, and therefore here we concentrate on precision (i.e. percentage of domain specific and syntactically well formed terms returned by the system).

In the case of SST, we are also interested in evaluating whether the ontological type of the term has been correctly recognized or not. Another relevant parameter to be taken into account is the percentage of terms which actually refer to concepts, since terms could refer to both concepts and instances, the former being the more interesting in building conceptualizations from scratch, the latter for the ontology population step.

For each term, the domain expert is asked to answer the following questions:

1) **WellFormed:** Is the term syntactically well formed? Is it meaningful? If yes mark Y in the corresponding filed, else mark N.
2) **Ontotype:** is the ontological type (i.e. supersense) assigned to the term correct? In other words, is the term reported in the correct list? If yes mark Y in the corresponding filed, else mark N.
3) **Pertinence:** is the term domain specific? Should it be included in the ontology? If yes mark Y in the corresponding filed, else mark N.
4) **Instance:** is the term an instance or a concept? If it is an instance mark Y in the corresponding filed, else mark N.

The four criteria above are exemplified in the following annotated table for the ontotype ANIMAL.

Term	**WellFormed**	**Ontotype**	**Pertinence**	**Instance**
Frequency over 1000				
Animal	Y	Y	Y	N
Elephant	Y	Y	N	N
Fish	Y	Y	Y	N
Frequency over 100				
Dogfish	Y	Y	Y	Y
Catfish	Y	Y	Y	Y
Ce	N	N	N	N
fishery_subsidy	Y	N	Y	-

In the example, the term *animal* is well formed, its ontological type has been properly recognized by the system, it is relevant for the domain, and it does not denote an instance. On the other hand, the term *elephant* is both well formed and correctly recognized as an animal, but it is not relevant for the fisheries domain. Terms like *dogfish* and *catfish* are instances of the class fish, therefore they are well formed, correctly recognized, pertinent to the domain and are instances. The term *ce* is simply a recognition error, therefore none of the properties above can be claimed for it. In general, all fields corresponding to improperly recognized terms should be leaved blank.

The case of *fishery_subsidy* is more complex. In fact it is a term, but its ontological type has not been properly recognized (it is not an animal). It is relevant for the domain and it is not clear

whether it should be treated either as an instance or as a class. In general the distinction between classes and instances depends on arbitrary decisions due to the ontology design requirements, therefore judging such kind of differences without the availability of an existing ontology is not that meaningful.

Data have been submitted to domain experts in an excel format, containing different terminological lists for different ontological types in different worksheets. The ontological types are indicated in at the bottom of the workbook, as illustrated below.

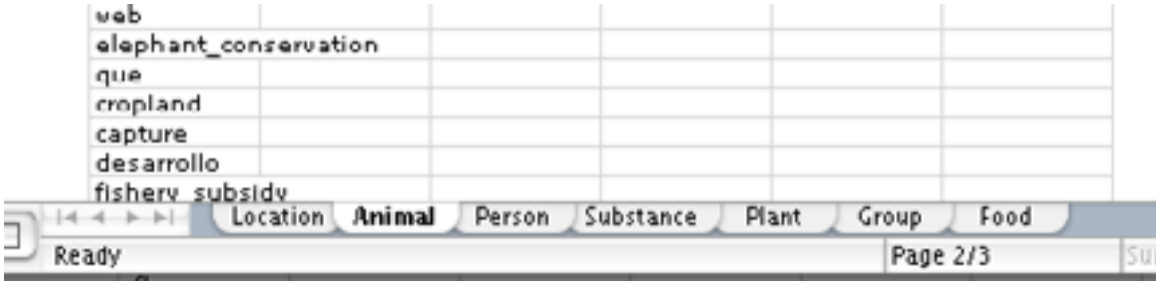

Figure 10. Screenshot of the evaluation interface adopted for the terminology extraction experiments

Domain Experts have been asked to fill each field on all tables, and to provide a qualitative report discussing the usability of such techniques, their impression and some suggestions for integration in the ontology engineering lifecycle. For evaluating the term extractor system, they have only been asked to compile the *well formed* and *pertinence* fields.

4.1.5 Results

Table 2 reports the compared evaluation we did on term extractor and SST. Results clearly shows that most of terms are well formed (i.e. they are correctly conjugated noun phrases) in both cases (around 85%). On the global list of returned terms, TermExtractor performed much better as far as the relevance with respect to the domain is concerned (pertinence), except for the category location.

Table 2. Compared evaluation of terminology extraction (Term Extractor vs SST)

	Well Formed	Ontotype	Pertinence	Instance
Term Extractor	0,85	NA	0,28	NA
SST-Food	0,94	0,59	0,18	0,49
SST-Plant	0,89	0,59	0,15	0,89
SST-Substance	0,81	0,32	0,12	0,27
SST-Animal	0,73	0,43	0,24	0,22
SST-Location	0,94	0,68	0,94	0,90

SST achieved bad results appended because we asked the lexicographer to validate samples form the full list of terms returned by SST, which includes also singletons. On the other hand, analyzing the top ranked part, the global figures increases sensibly, as shown in Table 3. In these settings, SST is much more accurate than term extractor.

Table 3. Evaluation of SST for terms with frequency above 1000

SST	WellFormed	Ontotype	Pertinence	Instance
Food	1	0,625	0,25	0,75
Plant	1	1	0,4	0
Substance	1	1	0,4	0,6
Animal	1	0,75	0,375	0,125
Location	1	0,8	0,85	0,4

Performances of term extractor and SST get closer as far as middle frequency terms are concerned (see Table 4).

Table 4. Evaluation of SST for terms with frequency between 100 and 1000

SST	Syntax	Ontotype	Pertinence	Instance
Food	1	0,7	0,2	0,733333333
Plant	0,95	0,578947368	0,111111111	0,166666667
Substance	1	0,45	0,25	0,5
Animal	0,866666667	0,576923077	0,230769231	0,384615385
Location	1	0,8	0,866666667	0,966666667

4.1.5.1 Effort required

4.1.5.1.1 Effort in input

Both methods we experimented for terminology extraction require a free text domain corpus as an input. No additional effort for domain adaptation is required in both cases, therefore they can be applied to any domain, provided that enough domain specific texts are available.

In addition, traditional terminology extraction techniques such as term extractor can be easily ported to different domains, since they only requires the availability of a Part of Speech Tagger and the rewriting of simple morpho-syntactic rules. In addition, systems for terminology extraction are freely available for most of the language. The same is not true for supervised technologies such as SST. In fact, to be ported to other domain it requires the availability of a large sense tagged corpus in the new language, where senses should be mapped to their corresponding supersense. Therefore, language adaptation is problematic, since the availability of such resources is very limited. In a recent work, SST has been successfully ported to Italian (Picca et al., forthcoming) by adopting MultiSemCor (Bentivogli et al, 2004) as a sense tagged corpus. Anyhow, the porting decreased the tagging performances due to the low quality of the Italian sense tagged corpus adopted for training.

4.1.5.1.2 Effort in output

As far as pertinence is concerned (i.e. the capability of extracting terms which are relevant for the domain of interest) both term extractor and SST obtained weak results (significant lower than 50% in both cases). It means that the effort required for validation is very high, since most of the terms, although syntactically well formed, should be checked first and then discarded. Performances of SST increases sensibly when concentrating on high frequency terms, therefore requiring a lower effort in validation as far as we are interested in building the high level parts of the taxonomy.

4.2 Similarity induction

The goal of this experiment is to help the domain expert in finding concepts/entities referring either to the same concept/entity or to a taxonomically related one. We approached this task by adopting distributional similarity techniques whose goal is to find, given a query instance/concept, similar instances/concepts of the same ontological type.

4.2.1 Use case

The domain expert is interested in enriching an already existing taxonomy of concepts, for example by including subclasses, hypernyms, co-hyponyms or synonyms of a given concept. To this aim he formulates a query to the system by simply selecting the concept/instance of interest in the ontology. As an output, the system provides the ranked list containing concepts/entities of the same type (e.g. if the query term is of type fishclass, the system retrieves only terms of type fishclass) found in the corpus and it submit them to the domain expert in a ranked list. Hopefully, synonyms and paradigmatically related terms are expected be placed in the top of the list.

4.2.2 Description of the technique

Our technique is based on the concept of semantic domain, introduced by Magnini et al (2001) and further explored by Gliozzo (2005). Semantic domains are common areas of human discussion, which demonstrate lexical coherence, such as *Economics*, *Politics*, *Law*, *Science*. At the *lexical level*, semantic domains identify clusters of (domain) related lexical-concepts, i.e. sets of highly paradigmatically related words also known as Semantic Fields.

Semantic domains can be described by Domain Models (DMs) (Gliozzo, 2005). A DM is a computational model for semantic domains that represents domain information at the term level, by defining a set of term clusters. Each cluster represents a Semantic Domain, i.e. a set of terms that often co-occur in texts having similar topics.

A DM is represented by a *k x k′* rectangular matrix **D**, containing the domain relevance for each term with respect to each domain, as illustrated in the following table.

Table 5. Example of Domain Model

	MEDICINE	COMPUTER_SCIENCE
HIV	1	0
AIDS	1	0
virus	0.5	0.5
laptop	0	1

DMs can be acquired from texts in a completely unsupervised way by exploiting a lexical coherence assumption. To this end, term clustering algorithms can be used with each cluster representing a Semantic Domain. The degree of association among terms and clusters, estimated by the learning algorithm, provides a domain relevance function. For our experiments we adopted a clustering strategy based on Deerwester (1990), following the methodology described in (Gliozzo, 2005). The input of the LSA process is a term-by-document matrix **T** reporting the term frequencies in the whole corpus for each term. The matrix is decomposed by means of a Singular Value Decomposition (SVD), identifying the principal components of **T**. This operation is done off-line, and can be efficiently performed on large corpora. SVD decomposes **T** into three matrixes $T \cong V\Sigma_{k'}U^T$ where $\Sigma_{k'}$ is the diagonal *k* x *k* matrix containing the highest *k′* << *k* eigenvalues of **T**

on the diagonal, and all the remaining elements are 0. The parameter *k'* is the dimensionality of the domain and can be fixed in advance[2]. Under this setting we define the domain matrix **D**[3] as $\mathbf{D_{LSA}} = \mathbf{I^N V}\sqrt{\mathbf{\Sigma_{k'}}}$.

Once the domain model has been acquired from the corpus analysis described above, it can be used to estimate the similarity between terms and documents in the domain space. In this space, the term t_i is represented by the i^{th} row of D, while documents are represented by the linear combination of the terms they contain. The similarity is then estimated by means of the cosine operation in this space.

For the purposes of similarity induction, it is possible to adopt the so generated space as follows. When a query Q is formulated (e.g. MUSIC), our algorithm retrieves the ranked list $dom(Q) = (t_1 , t_2 , \ldots , t_k)$ of domain specific terms such that $sim(t_i , Q) > \theta$ where $sim(Q, t)$ is the cosine between the DVs corresponding to Q and t, capturing domain proximity, and θ_t is the domain specificity threshold. The process is illustrated in Figure 11. The output of the similarity induction step is then a ranked list of similar terms.

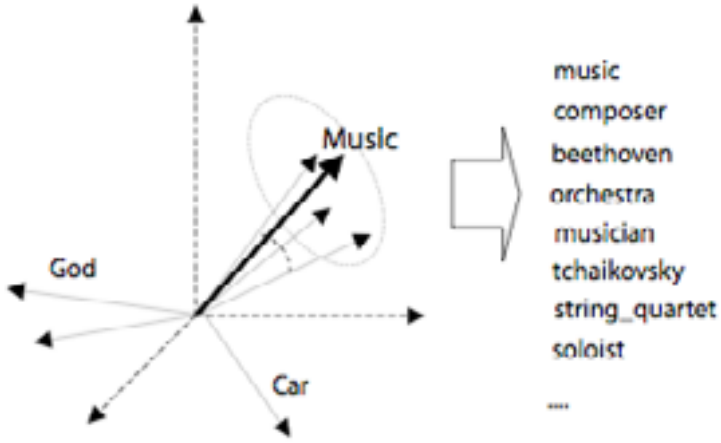

Figure 11. Semantic Domain generated by the query MUSIC

Applied to our experimental settings, we adapted the standard domain modelling technique to the purposes of similarity induction by indexing typed concept/instances instead of terms in the term by document matrix, and then running the SVD process on the so obtained matrix. Entities and concepts have been recognized in the FAO corporate document repository by adopting GATE, as described in Chapter 3. By following this methodology it is possible to retrieve concepts/entities of the desidered type by simply specifying this constraint. Example of output are reported below

Query: FISHCLASS Pandalus_borealis:

"FISHCLASS#Pandalus_borealis"

"FISHCLASS#Pandalus_montagui"

"FISHCLASS#Crangon_crangon"

"FISHCLASS#Pandalus_jordani"

[2] It is not clear how to choose the right dimensionality. In our experiments we used 100 dimensions.

[3] Details of this operation can be found in (Gliozzo, 2005).

"FISHCLASS#Penaeus_schmitti"

"FISHCLASS#Penaeus_setiferus"

4.2.3 Evaluation

To evaluate the similarity induction experiments we performed a simple experiment consisting on estimating the probability of finding taxonomically related terms at different ranges of similarity. To this aim, we randomly sampled a set of 33 instances of type fish_class, and we submitted 33 queries to the similarity induction system, requiring as an output instances of the same type having a similarity above 0.6. Overall, we obtained 243 candidate related pairs of terms.

Since by construction all those terms refers to concepts in the ontology (see the annotation process adopted for NER in section 3), we where able to automatically evaluate our algorithm by exploiting the taxonomical structure of the reference ontology. In particular, for each concept, which was returned by the system, we calculated the most specific ancestor in common with the query concept in the ontology, and we evaluated the capability of our algorithm of finding such similarities at different level of abstraction. The idea is that the more specific the ancestor, the better the similarity induction process. Therefore we categorized each candidate pair onto the following categories:

- **Concept:** fish1 and fish2 are synonyms, i.e. both lexicalize the same concept,
- **Genus:** fish1 and fish2 share the same genus
- **Family:** fish1 and fish2 share the same family
- **Order**: fish1 and fish2 share the same order
- **Major group:** fish1 and fish2 share the same major group (i.e. they are both pisces)
- **Unrelated:** if fish1 and fish2 are unrelated

Then we computed the probability of finding the relations above between the query term and the retrieved terms at different ranges of similarity. Results are reported in Figure 12.

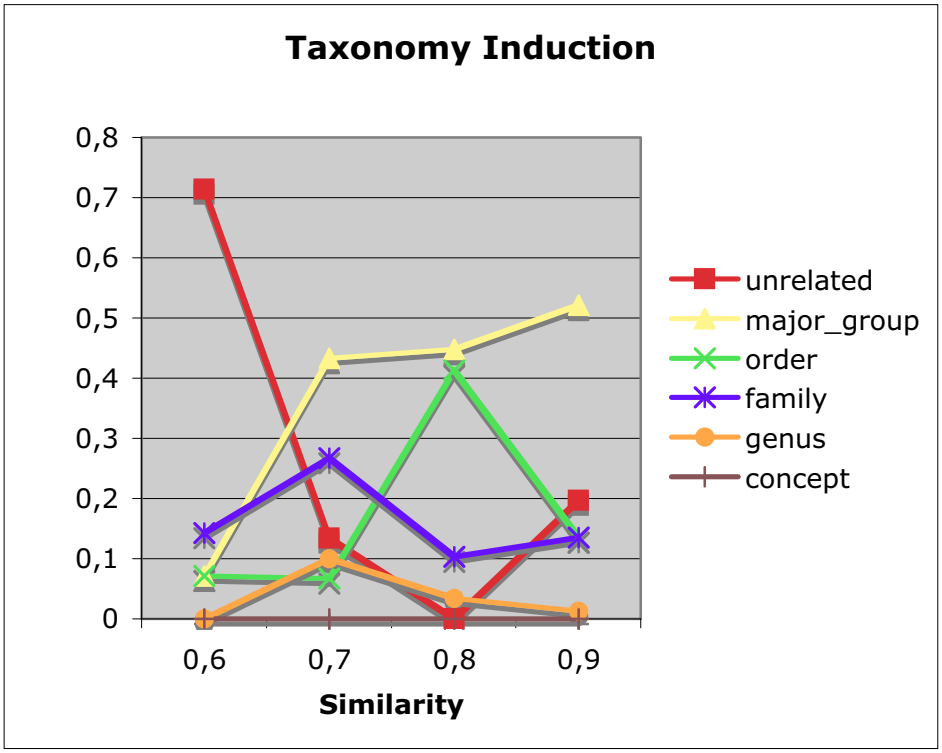

Figure 12. Probability of finding taxonomically related terms at different similarity ranges

Evaluation shows that there exists a strong correlation between the probability of finding taxonomic relations among terms and their similarity. Even if distributional similarity approaches are not capable of identifying strict relations such as synonymy, the probability of finding instances belonging to the same order or mayor group is around 0.5 for higher ranked similarity pairs. In addition, the probability of finding unrelated terms goes to zero when the domain similarity between the query and the returned terms increases. This is a very nice property, since it provides an effective methodology to assist the ontology editor in structuring the ontology and finding potential candidate related concepts/instances, for example by filtering out the potentially non-related terms.

4.3 Relation extraction

For the Relation Extraction experiments we adopted two different approaches on two different corpora: pattern based approaches have been adopted to acquire information from Factsheets, while distributional approaches have been preferred to extract information from natural language texts. The reason of this choice is that fact sheets allowed us to easily identify highly accurate patterns in texts, mostly based on the XML document structure. On the other hand, identifying specific patterns on the FAO document collection was more difficult, since the language variability prevents us from achieving high performances. In the following two subsections we will describe the basics of both approaches and the achieved results.

4.3.1 LSA based approach

4.3.1.1 Use case

The Domain Expert is adding a new entity/concept in the ontology. She/He is interested in the relations between the new entity and the remaining entities of a specific type in the ontology (e.g.

the water body where a particular species of fish lives). She/He types a query to the LSA system, by simply indicating a concept/entity already in the ontology. As an output, the system retrieves the ranked list of concepts/entities already in the ontology, where the ranking is estimated according to their LSA similarity computed on a domain specific corpus. The system also returns, for each instance/concept, the links to their occurrences in the corpus (e.g. providing snippets which can be further expanded in analogy to what standard search engines do). Going through the ranked list, the domain expert judge whether the retrieved instance/concept is related to the query instances/concepts according to the semantic relation of interest (e.g. lives in), and eventually ask the system to include the new relation in the ontology.

4.3.1.2 Description of the technique

The relation extraction technique here proposed is based on the work described in (Gliozzo et al., 2007), consisting on ranking candidate pairs of related entities according to their LSA similarity. In our settings, the LSA similarity is estimated by performing a Singular Valued Decomposition on the entity by document matrix obtained by recognizing occurrences of both Entities and Concepts in the corpus. Such occurrences have been identified by adopting the NER technique implemented in GATE based on dictionary lookup described in Chapter 3. In particular, from the FAO corporate document repository we extracted an entity by document matrix, and we performed the SVD process on that. To this aim we exploited the domain modelling techniques described in Section 4.2.2. Then, for each entity of type Large_Marine_Ecosystem we ranked all the entities of the class Fish_Class according to their LSA similarity. An example of the output is provided below.

The hypothesis at the basis of the exploitation of LSA based techniques for relation extraction is that in all those cases where there exists a “privileged” relation between entities of specified ontological types, the similarity between those is in itself a reliable estimator for the probability that such a relation holds. For example, the relation between the concepts fish_species and

Bay_of_Bengal	Parastromateus_niger	0,99893713
Bay_of_Bengal	Atropus_atropos	0,99893713
Bay_of_Bengal	Polynemus_paradiseus	0,99893713
Bay_of_Bengal	Lepturacanthus_savala	0,99893713
Bay_of_Bengal	Scomberomorus_guttatus	0,99885094
Bay_of_Bengal	Selaroides_leptolepis	0,99793124
Bay_of_Bengal	Trichiurus_lepturus	0,9978997
Bay_of_Bengal	Polydactylus_sextarius	0,9978997

large_marine_ecosystems in the fisheries domain is almost always the **lives_in**. “Privileged relations” can be discovered very often in most of domain ontologies, as the highly constrained domains they represent often allow the existence of a very limited set of relations among entity pairs of specified types. Another example is the relation Person:X work_for Organization:Y which very often holds in ontologies describing organizations. Therefore, our methodology is quite general and can be generalized to any domain without modifications.

4.3.1.3 Evaluation

4.3.1.3.1 Experimental settings

To evaluate our method we exploited the FAO corporate document repository where all entities belonging to Large_marine_ecosystem and Fish_species have been previously recognized, and we performed LSA on the so obtained entity by document matrix. For each entity belonging to the former class, we ranked all the entities belonging to the Fish_species class according to their LSA similarity, selecting only those entity pairs having a similarity higher than 0.5. Overall, we extracted more than 3000 candidate relations. Out of those, we randomly selected a subset of 25 queries

(i.e. entities of the class large_marine_ecosystem) , and we submitted the 10 top ranked instances for each query (i.e. the top 10 stronger associated entities of the class fish_species) to the domain experts by adopting the validation interface presented in Section 3.3.1. Guidelines for evaluation have been submitted to the evaluators. Overall, the domain experts validated 246 pairs of instances. Table 6 reports the number of extracted relations obtained by our method by setting different similarity thresholds.

Table 6. Number of extracted candidate related pairs

Min similarity	# extracted relations
0,9	74
0,8	259
0,7	362

4.3.1.3.2 Results

Results of the evaluation are summarized in Figure 13, which reports the probability of finding correct relations (i.e. precision) at different similarity ranges. They clearly show that the LSA similarity between candidate related pairs is strongly correlated with the precision of the algorithm, which reaches 0.75 when the similarity is above 0.9. Quite acceptable results (i.e. precision close to 0.6) are also achieved when the similarity is above 0.7. The same figure reports the percentage of relations that can be found for any similarity range with respect to the overall number of extracted relations. It shows that, even if most of the pairs have a similarity below 0.6 (i.e. they are not related), around 10% of them are actually very likely to be related.

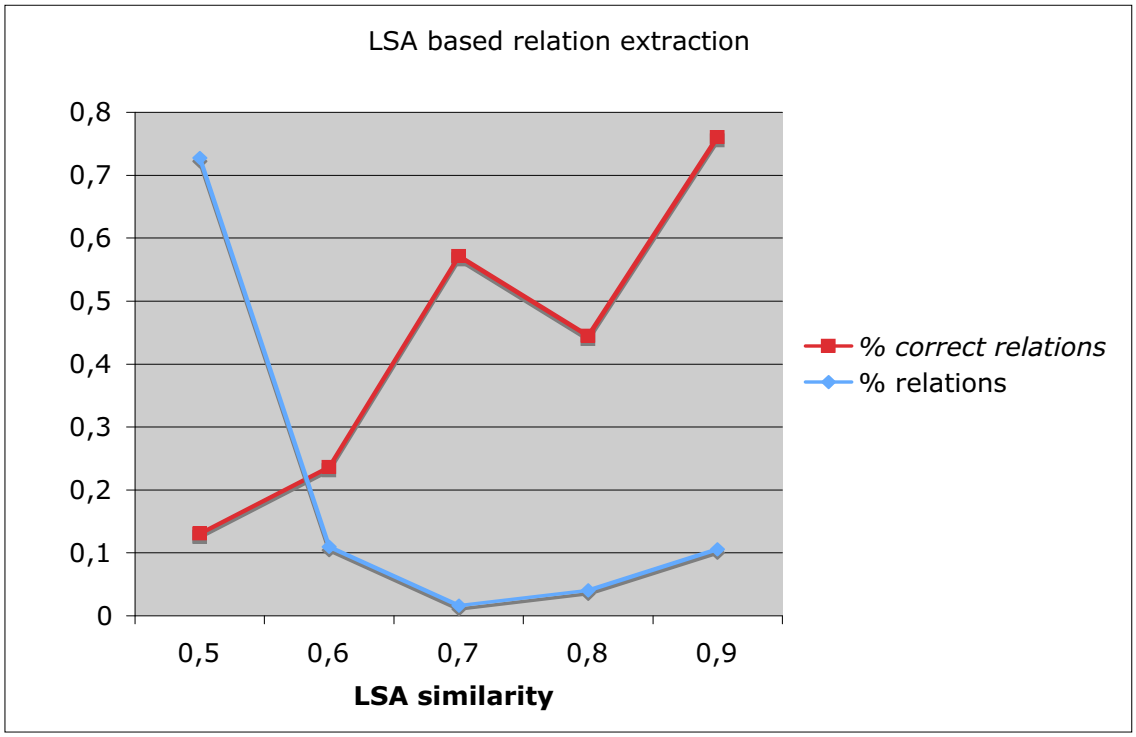

Figure 13. Precision versus LSA similarity

Table 7 reports an estimation of the total number of acquired relations on the overall dataset (i.e. including also all those relations that have not been manually checked) that can be found at different level of expected accuracy. It shows that the algorithm is robust and accurate enough to ensure a nice recall, since around 220 correct *lives_in* relations can be extracted from a sample of 362 relations having an accuracy of 0.63.

Table 7. Total Number of extracted relations at different similarity ranges

Similarity Range	Expected accuracy	Number of relations
> 0.9	0.76	74
> 0.8	0.62	259
> 0.7	0.63	362

4.3.1.4 Effort required

4.3.1.4.1 Effort in input

The LSA-based relation extraction algorithm is totally general and can be applied to any domain, ontology, relation and language without requiring additional costs for domain adaptation. In fact, it is based on a NER procedure based on dictionary lookup, which only requires a specific morpho-syntactic analysis for any different language, and can be applied uniformly for different domains. Since this operation can be easily performed on most of the natural languages with existing pre-processing tools such as GATE, it can be easily ported to any language. The LSA algorithm is totally independent from the language, since it is based on a linear algebraic operation that can be applied to any rectangular matrix. Finally, once the LSA based representation of all entities in the ontology is performed, the relation extraction process is simply a query in the LSA space, that can be performed by using any entity in the ontology, therefore it can be applied to any domain, language, ontology and relation of interest. In a scenario in which the overall learning process is fully automatic, the only effort required for input is the selection of both a domain specific corpus and a reference ontology to be further populated. Therefore, we conclude that the algorithm basically do not require any effort in input.

4.3.1.4.2 Effort in Validation

The effort required for validation is proportional to the percentage of correct relations that can be found as an output of the query. In fact, considering as a constant the time required to check the validity of any pair, to collect a certain amount *n* of valid relations the domain expert needs to validate *n* x *p* relations, where *p* is the expected accuracy of the extraction algorithm. Assuming that the validation of each pair requires 5 minutes, we estimate that by using the LSA method it is possible to acquire around 220 correct relations (i.e. the 60% correct relations out of the 362 relations having similarity above 0.7) in around 3 days of man work. These figures are clearly overestimated, since relations are submitted for validation as a ranked list of entities related to the same query. It clearly simplifies the validation process.

5.3.1 Pattern-based Relation Extraction

A lot of information about species, fishing gears and vessels is not yet contained in the RTMS or any of the ontologies, but still hidden in semi-structured FIGIS factsheets. As a natural first step towards populating the fishery ontologies with instantiations of particular relations, we therefore implemented a relatively simple relation extraction tool that exploits the structure of the factsheets. In the following we describe our approach in more detail and present the results of an evaluation that was carried out at FAO.

5.3.1.1 Approach and Implementation

In a first series of experiments, we extracted three types of relations from species factsheets: <*species*> *has-synonym* <*string*>, <*species*> *lives-in* <*location*> and <*species*> *lives-not-in* <*location*>. For all these relations there is a dedicated paragraph in each factsheet that contains unstructured or semi-structured text and that is easily recognizable by its caption ("Synonyms" or "Geographical Distribution", respectively). Once we have identified the relevant paragraph we can use shallow parsing techniques and lexico-syntactic patterns to extract from it all the noun phrases that are likely to fill the range of the relation.

For example, consider the factsheet for 'Caretta Caretta', also known as 'Loggerhead Sea Turtle'[4]. The "Synonyms" paragraph contains a list of synonyms and associated references to the literature.

Synonyms

* *Testudo Cephalo* Schneider, 1783

* *Testudo Caouana* Lacepède, 1788

* *Chelone caretta* Brongniart, 1805

* *Chelonia Caouanna* Schweigger, 1812

* *Caretta nasuta* Rafinesque, 1814

* *Chelonia cavanna* Oken, 1816

* *Caretta atra* Merrem, 1820

[...]

The extraction of all synonyms from this paragraph is straightforward thanks to the fact that the synonyms, unlike the names of their inventor or the year of their first publication, are written in italic style. Thus, we can easily generate the following list, each item of which corresponds to an instantiation of the has-synonym property for the individual Caretta_Caretta.

[Testudo Cephalo, Testudo Caouana, Chelone caretta, Chelonia Caouanna, Caretta nasuta, Chelonia cavanna, Caretta atra, Caretta Cephalo, Caretta nasicornis, Chelonia caretta, Testudo Corianna, Chelonia pelasgorum, Chelonia cephalo, Chelonia (Caretta) cephalo, Chelonia caouana, Chelonia (Thallassochelys) Caouana, Chelonia (Thalassochelys) atra, Thalassochelys caretta, Chelonia (Caouana) cephalo, Halichelys atra, Caouana Caretta, Caouana elongata, Thalassochelys Caouana, Thalassochelys corticata, Chelonia corticata, Thalassochelys elongata, Thalassiochelis caouana, Eremonia elongata, Caretta caretta, Caretta caretta, Thallasochelys cephalo, Caretta caretta caretta, Caretta gigas, Caretta caretta gigas, Caretta caretta tarapacana]

More difficult seems the extraction of range fillers from the paragraph that describes the geographical distribution of Caretta Caretta.

Geographical Distribution

Caretta caretta is widely distributed in coastal tropical and subtropical waters (16-20°C) around the world. Commonly this species wanders into temperate waters and to the boundaries of warm currents. It is suspected that some loggerhead turtles undertake long migrations using warm currents (e.g., the Gulf Stream in the North Atlantic; the North Equatorial and Kuroshio Currents and the California Current (12-20°C) in the North Pacific and other currents in the southern hemisphere), that

[4] http://www.fao.org/fi/website/FIRetrieveAction.do?dom=species&fid=2748

bring them far from the nesting and feeding grounds. Apparently, the limit of distribution is waters of about 10°C; if they encounter colder waters, they may become stunned, drift helplessly and strand on nearby shores. Records are quoted from New England and eastern Canada, Labrador and Nova Scotia, especially between July and October of warm years. The northern limit of distribution is a summer capture of a live young turtle entangled in a fishing line off Murmansk, Barents Sea (68° SS'N). Brongersma (1972) quotes this and many other records for European waters. Occasionally, the species is sighted in southern Australia and New Zealand. In South America it is absent from west Colombia, Ecuador and Peru, but there are some records from Arica and Coquimbo, in Chile; on the eastern coast, the southernmost record is Rio de la Plata, Argentina.

As for all the other factsheets we analysed, most of the noun phrases contained in this paragraph represent locations (countries or water bodies) that are inhabited by the species – in this case, Caretta Caretta. However, there are a few exceptions. On the one hand, there are proper nouns such as 'July' or 'October' that indicate the times of the year when the species occurs in certain regions. These could be filtered out by a named entity classifier (this is left for future work) that distinguishes, e.g., between dates and locations. On the other hand, there are locations which Caretta Caretta is said to be absent from such as 'Colombia', 'Ecuador' and 'Chile'. We identified this type of negative information by searching for particular keywords, including 'not', 'never', 'except', 'absent' or 'excluding'.

Finally, our approach would extract from the above cited paragraph, e.g., the following individuals as potentials ranges for the *lives-in* relation: 'New England', 'eastern Canada', 'Nova Scotia', ' Murmansk', 'Brongersma', 'southern Australia', 'New Zealand', 'South America', 'Arica, Coquimbo', 'Chile', 'Rio de la Plata', 'Argentina'.

5.3.1.2 Experiments and Evaluation

In order to evaluate our approach we analysed an overall number of 380 species factsheets in HTML format. From these documents we obtained 313 sets of synonyms[5], 3,160 instances of the *lives-in* relation and 10 instantiations of *lives-not-in*. A randomly chosen subset of these results was given to an FAO fishery expert[6] who used the evaluation described in section 3.3.1 to assess the quality of each *lives-in* and *has-synonyms* relation. A similar form was provided for instantiations of the *lives-not-in* relation. Table 8 gives an overview of the evaluation results, in particular the accuracy (#correct/#relations) we obtained for the various kinds of relationships.

Table 8. Evaluation of the patter based relation extraction algorithm

	# Relations	# Correct	Accuracy
has-synonyms	34	26	0.76
lives-in	240	209	0.87
lives-not-in	10	6	0.60

[5] We evaluated sets of synonyms instead of presenting each synonym separately to the expert. The reason is that we wanted to reduce the effort required for the human evaluation and focus on the lives-in relation instead, which we assumed to be more difficult to extract.

[6] Aureliano Gentile, FIES

In addition to the overall assessment the human expert provided us with valuable feedback and comments some of which are reported in the following.

- Some location names are only partially correct. ***Example:*** 'Atlantic Coast of Middle', 'Western Atlantic : Greater' ***Explanation:*** This is a bug in the noun phrase extraction rule. Apparently, the colon is not recognized as a token delimiter.
- Several locations that are actually mentioned in the text were not extracted. ***Explanation:*** In some cases, this is true and can be explained by incorrect part-of-speech tags or tokenizer errors. However, most of the location names that are said to be missing here, were actually extracted, but not shown to the human expert. As pointed out before, we had to choose a subset of our results for the evaluation in order to reduce the effort of manual assessment.
- The synonym extraction sometimes erroneously identifies latin expressions like 'et al.' as species names, or the extracted synonyms are incomplete. ***Example:*** '[Clupanodon jussieu, nomen dubium, Spratella tembang, Clupea immaculata, Sardinia immaculata, Fimbriclupea dactylolepis, Sardinella dactylolepis, Sardinella taiwanensis, Sardinella jussieu, Sardinella jussieui, et al, gibbosa, tembang, jussieu, Sardinella gibbosa]' ***Explanation:*** This error is caused by the fact that the synonym extraction pattern does not rely on linguistic information, but only considers the font style in order to identify potential synonyms in the dedicated paragraph. This heuristic seems to fail, e.g., in the following cases: "*Sardina sagax* , , (part):Regan, 1916:13 (combined with *sagax*, *ocellata* and *caerulea*).", "*Sardinops sagax melanosticta* Svetovidov, 1952:178, pl. 6, fig. 1; *Idem*, 1963:193, pl. 6, fig. 1"
- In some cases, the scientific name that is used as the domain for each relation instantiation seems to be incorrect with respect to the associated synonyms. ***Example:*** 'Scomber thynnus', 'Dussumieria elopsoides' ***Explanation:*** These might be inconsistencies in the factsheets rather than extraction errors.
- The token 'S' is often mistaken for a location name. ***Example:*** 'Thunnus alalunga *lives-in* S' ***Explanation:*** In the factsheets, 'S', 'N', 'W' and 'E' are frequently used to specify directions. Since they are spelled with upper-case letters the part-of-speech tagger considers them as nouns that is potential candidates for location names.

5.3.1.3 Discussion

This overall approach is simple, but efficient for a limited set of relations. It does not require a deep linguistic analysis and can be implemented in a way that is scales up to thousands or hundred thousands of documents. The patterns for identifying relevant paragraphs are relatively simple regular expressions and can be written without any linguistic knowledge. Even the patterns for detecting noun phrases in larger snippets of unstructured text are not too complicated and general enough to be reused for other kinds of relations.

Despite all these advantages, the current implementation of our approach still has a number of drawbacks. First, all patterns for identifying relevant paragraphs and noun phrases are hard-wired in the code and specified by means of regular expressions over markup, tokens and their syntactic categories. In order to facilitate maintenance and adaptation of these patterns, a declarative specification would be preferable. Second, at least the patterns for identifying relevant passages of text are very specific with respect to the structure of the documents. Thus, every change to the underlying XML schema would require a modification of the code. Approaches to wrapper induction that are a popular means for information extraction from web pages could facilitate the necessary adaptation process in this case. Moreover, automatic pattern induction approaches (e.g., *Espresso*, *KnowItAll*) might be useful for discovering new patterns, even for relations that are not reflected in the coarse-grained structure of the factsheets. However, this kind of approaches typically yield a relatively bad precision compared to methods that rely on manually specified extraction rules.

4.4 Ontology Mapping

4.4.1 Use Case

A domain expert has to use jointly two knowledge structures (ontologies, thesauri) and therefore he is interested in the relations that hold between the entities of these structures. These relations can be equivalences, subsumption relations or disjointness (incompatibility) relations. A matching system is used to produce these relations in the form of an alignment, i.e. a set of mappings between the concepts of the input knowledge structures. Each mapping contains the source and target concepts, the relation that the system derived between them (i.e., one of equivalent, subclass, superclass, disjoint), an ontology (or a set of ontologies) that was used to deduced the relation, as well as the inference steps that lead to the relation.

4.4.2 Description of the Technique

Our ontology matching tool, SCARLET (Semantic Relation Discovery by Harvesting Online Ontologies), is based on the idea of using the entire Semantic Web as a source of background knowledge for ontology matching. As we have described in (Sabou et al,2006), our matcher automatically finds and explores multiple and heterogeneous online knowledge sources. For example, when matching two concepts labelled *Researcher* and *AcademicStaff*, our matcher 1) identifies (during matching) online ontologies that can provide information about how these two concepts inter-relate and then 2) combines this information to infer the mapping relation. The mapping can be either provided by a single ontology (e.g., stating that a *Researcher* isA *AcademicStaff*), or by reasoning over information spread in several ontologies (e.g., that *Researcher* isA *ResearchStaff* in one ontology and that *ResearchStaff* isA *AcademicStaff* in another). A full description of this matcher is provided in D3.4.1 (NeonD3.2.4).

A prototype matcher based on this technique has been implemented on top of the Swoogle Semantic Search Engine, and used to perform the matching experiments reported here. Note that an extended version of this prototype is currently being implemented in conjunction with the WATSON semantic web gateway (d'Aquin et al, 2007).

4.4.3 Evaluation

4.4.3.1 Experimental Setup

We have performed two major matching experiments, described in the following paragraphs:

4.4.3.1.1 Mapping AGROVOC and NALT

In the first experiment, we matched the AGROVOC and NAL thesauri. The United Nations Food and Agriculture Organization (FAO)'s **AGROVOC** thesaurus, version May 2006, consists of 28.174 descriptor terms (i.e., preferred terms) and 10.028 non-descriptor terms (i.e., alternative terms). The United States National Agricultural Library (NAL) Agricultural thesaurus **NALT**, version 2006, consists of 41.577 descriptor terms and 24.525 non-descriptor terms. Since alternate terms often describe synonyms of the preferred terms, we have also used those in our experiments. Therefore, for each concept we consider all its labels (descriptors). In the case of AGROVOC we only relied on English labels. Note that these thesauri describe a broad range of domains, from animal species to chemical substances and information technology. The matching process lead to a total of 6687 mappings containing 2330 subclass, 3710 superclass and 647 disjoint relations.

For evaluation purposes, we randomly selected 1000 mappings (i.e., 15% of the alignment) containing an appropriate proportion of different mapping relations, namely: 100 disjuncts, 350 subclass, 550 superclass relations. As evaluators, we relied on six members of our lab working in the area of the Semantic Web, and thus familiar with ontologies and ontology modeling. We performed two parallel evaluations of the sample mappings (i.e., each mapping has been evaluated by two different evaluators). The participants were asked to evaluate each mapping as Correct, False or "I don't know" for cases where they could not judge the correctness of the statement. The participants were allowed to use any kind of material (e.g., (web-)dictionaries, Google) in cases where they were not familiar with the domain and needed some more information for evaluating a given mapping (e.g., when judging that *Leukemia* isA *Neoplasm*). A specialized graphical interface (see Section 3.3) has been developed to facilitate the task of the evaluators by displaying the mappings together with the context in which the mapped concepts appeared in the source ontologies (i.e., their semantic neighborhood).

4.4.3.1.2 Mapping AGROVOC and ASFA

In the second experiment, we performed a matching between two ontologies provided by FAO, namely AGROVOC and ASFA. ASFA contains 9021 concepts covered by four major namespaces: asfa (6610), WordNet(1994), dolce(225) and FiCore(162). We performed a matching only between the 6610 concepts specific to ASFA. We obtained 3479 mappings: 1949 superclass relations, 1205 subclass relations, 285 disjoint relations and 40 equivalence relations. In this case, the evaluation was performed by two FAO domain experts. The first expert evaluated a sample of randomly selected 200 mappings while the second assessed 500 mappings. The experts were also supported in their evaluation by the graphical interface described in Section 3.3.

4.4.3.1.3 Results

In the case of the first experiment where AGROVOC and NALT were matched, we computed the precision of the obtained alignment based on the evaluation provided by both groups. Table 9 summarizes the number of Correct, False and unevaluated (Don't know) mappings for each group, as well as the number of these mappings agreed by both groups. The two groups agree on 742 mappings (we exclude the "Don't know" answers because there are no real agreements on those), and therefore have an agreement coefficient of 74%. We define the precision of the alignment as the ratio of Correct mappings over all the evaluated mappings (i.e., those evaluated either as Correct or False).

Table 9. Evaluation results for the AGROVOC - NALT matching

	Group 1	**Group 2**	**Agreed by All**
Correct	586	666	525
False	346	299	217
Don't know	68	35	10
Precision	**63%**	**69%**	**70%**

We obtained precision values of 63% and 69% for the two groups. The gap between these values is due to the variation in the way evaluators performed their task: some investigated each mapping thoroughly, while others simply provided no evaluation for the mappings they were not sure about. To level out these differences, we also computed the precision of the part of the alignment on which both groups agreed, as we think this better reflects the typical performance that can be achieved with our paradigm. In this case, the precision was equal to 70%.

The results obtained for the second experiment are shown in Table 10. Unlike in the case of the first experiment, the precision values obtained in this second experiment are much lower. We are

currently investigating the causes of these low performance values. Our main hypothesis concerning the relatively low performance of our matcher on AGROVOC and ASFA is that the overlap between these two resources may be too low to obtain significant results, and that the terms in ASFA may not be well covered by the Semantic Web. If, in further analyses, these hypotheses appear to be validated, it would mean that the actual performance of the matcher is highly dependent on the resources to match. A mechanism allowing the user to get *a priori* an indication of the performance of the technique could then be envisaged, distinguishing before matching if the use of our matcher is worth the effort (like in the case of AGROVOC and NALT), or if the use of a more appropriate matcher is advisable (like in the case of AGROVOC and ASFA).

Table 10. Evaluation results for the AGROVOC - ASFA matching.

	Expert 1 (200 mappings)	**Expert 2 (500 mappings)**
Correct	32	141
False	168	358
Don't know	0	1
Precision	**16%**	**28.5%**

4.4.4 Effort Required

4.4.4.1 Effort during Input

Unlike existing background knowledge based matchers where the appropriate background knowledge needs to be selected a priori (see Section 3.2), SCARLET automatically identifies appropriate bits of knowledge provided by online ontologies. As such, the matcher does not require anything else than the knowledge structures that need to be matched. Also, it is completely domain independent.

4.4.4.2 Effort during evaluation

Due to the general lack of gold standards in the matching field, evaluation is usually performed manually by one or more domain experts. We have tried to make this process as easy as possible by building a specialised graphical interface which presents the evaluator with each mapping (and the context of the matched concepts) and allows him to evaluate the mapping by simply clicking the appropriate result (i.e., correct, false, don't know).

5 Recommendations

This deliverable presented a set of ontology learning experiments we performed in the fisheries domain in the context of WP 7 of the NeOn project. The main contribution of this deliverable to the global picture of WP7 is the set of recommendations reported in this section, since they will lead further development and integration of the experimented technology in the NeOn toolkit. According to the original spirit of the deliverable, recommendations will be provided in term of pros and cons related to the exploitation of the selected technologies, with particular regards with respect to the reduction of effort allowed by the automatic technique for the ontology population and maintenance process. As introduced in Sec. 1.2.5, we identified three dimensions for evaluating each technique: effort required for domain adaptation (input), effort required for validation (output) and implementation costs for each technique. The effort required in input is directly related to the degree of generality of each technique (i.e. the more general, the more domain independent). The effort required in output is related to the accuracy of the system, since a low accuracy (let's say 0.2) will require an higher amount of manual checking to achieve the same amount of knowledge (in the example, to collect 20 correct relations, the domain expert is asked to validate 100 examples, while if the precision of the algorithm were 0.8, only 25 examples would be validated to collect the same amount of data). The implementation costs are related to the availability of open source code for the examined technique and to the difficulty of engineering the algorithms adopted.

Terminology extraction

Identifying a terminological list is the first step of any ontology design process, and it is crucial especially at the very early stages of domain ontology development. Since in many cases domain corpora are available and terminology induction systems are fully unsupervised (they do not require any effort in input), the exploitation of terminology extraction techniques should be recommended, even in spite of the low accuracy reported in our experiments. In fact, as far as the most frequent terms are concerned, terminology extraction systems provide the basic terms composing the top level part of the ontology with a reasonable precision (around 30%). Further, the exploitation of the SST technology provides typed terminological lists, identifying both concepts and entities referring to locations, persons, organizations, animals, and so on. Even if for ontology learning it is really important to have fine-grained classification into very many classes, in our experiments we simple selected coarser grained classes to avoid any requirement of training data for the finer grained cases, the development of which will increase enormously the effort in input. Adopting this strategy, the automatic learning task is too far removed from the actual practical challenge and should be integrated with a finer grained manual categorization step, which is anyhow required by adopting a standard terminology extraction technology.

Our experiments show that terminology extraction systems can be used to augment the coverage of the conceptualization, for example by highlighting new concepts which have not been included yet in the thesaurus adopted by the domain experts. On the other hand, state of the art technology for terminology extraction is still weak, and requires a large effort in validation due to the low accuracy of the returned results, at least as far as very specific domains (such as fishery, in our use case) are concerned. Therefore we recommend the exploitation of state of the art terminology induction (either term extractor or SST) to domain experts just to get a first idea of the concepts required to describe new domains when starting from a corpus, but we do not suggest to extensively check the full list of terms with the goal of adding the pertinent ones. In addition, we would like to remark the crucial importance of this technology, and to push the further development of innovative solution inside the NeOn project aiming on increasing the precision of the algorithms.

Similarity induction

Although the induction of taxonomies has not been highlighted as a priority from the user requirement perspective, similarity induction techniques are at the basis of ontology learning and ontology engineering, since they provide unsupervised methods to find analogies between concepts and instances in a very cheap and cost effective way. In addition, the development of highly accurate similarity induction methods is a prerequisite for the development of relation extraction algorithms. We experimented with the effectiveness of similarity induction in providing useful knowledge to build taxonomies, i.e. its capability of finding hypernyms, hyponyms or synonyms of a selected concept in the ontology. The experiments, performed in a controlled setting, in which evaluation has been performed against the reference ontology itself, which has a very coarse-grained classification, show that most of highly similar terms are also related somehow with some taxonomic relation, most of them belonging to the same major group or order and family. For our evaluation, very closely connected terms, such as synonyms, were discovered less frequently. Since the effort required in input is almost null, the additional effort required by adopting this method is very low, motivating the exploitation of this technique when high accuracy is not required,

Therefore, we do not recommend the use of similarity based techniques by themselves to solve taxonomy induction problems, but rather to use them in combination with other techniques. In addition, similarity based tools could be used by the domain expert to find useful suggestions during for ontology design, for example by querying a system whenever a concepts is not very well understood or conceptualized.

Relation Extraction

Regarding the relation extraction process, we experimented with two different approaches: pattern based and similarity based.

Pattern based techniques are well known and largely applied due to their simplicity ad efficiency. They are known to obtain the best results when applied to semi structured text, as opposed to free text, where such approaches are typically not well performing. We adopted the pattern based approach to the fisheries fact sheets, obtaining precision at around 0.85. This result suggests that this technique can be adopted to support editors in their work of adding relations to existing ontologies. However, since patterns need to be adapted manually, and not all domain experts can be assumed to be fluent in regular expressions, editors should be assisted in this task by appropriate user interfaces. A compromise solution would be the use of simplified interfaces (such as those integrated in GATE or Text2Onto) to write such expressions, or the exploitation of supervised learning techniques to acquire patterns from examples. The latter techniques, however, are generally much less accurate, since they can't be controlled and require a lot of examples. Regarding the implementation cost, pattern based approaches should be preferred since they are very simple and effective algorithms, which can be engineered in a few weeks to be adapted to any environment.

Alternatively, or in conjunction with user interface, one could provide editors with supervised learning technique to acquire patterns from examples. This strategy has not been followed during our experiments due to time constraints. However, in this latter case, a large amount of examples is needed, resulting in a considerable increase of the effort in input. As an alternative, there also exist approaches based on bootstrapping, which basically require examples of related pairs as an input instead of labelled texts (e.g. the ESPRESSO system (Pantel and Pennacchiotti, 2007)). In this case, the expected accuracy is lower, while the effort required in input is practically null.

On the other hand, similarity based approaches show complementary features with respect to the above. In fact, when used in combination with NER techniques, similarity based techniques demonstrated a good capability of finding related terms with quite high precision (above 0.5, in our experiments) on free texts, overcoming the limitations of pattern based approaches which are more effective when applied to semi-structured text. In summary, similarity based approaches do not

require any effort in input at the cost of some effort on validation, required to filter out irrelevant relations.

Given these results, our recommendation is to further exploit the possibilities of pattern based approaches applied to semi-structured texts. In particular, special attention should be paid to the graphical interface to make available to the experts and to the interaction model suitable for inclusion in the lifecycle. In addition, we recommend the use of similarity based approaches when working on free texts, such as documents collections and websites, and in particular when the domain is not strictly defined. In this case, the effort in validation would be higher due to the lower accuracy, while the effort in input will be practically null. Both approaches should be further developed to be integrated in the neon toolkit and can be applied to any language and domain.

Ontology mappings

The task of ontology mapping has been identified as a priority from the user requirement analysis, since it is a crucial step for knowledge integration and interoperability. Therefore finding an affective approach for this task is crucial to fit the user requirements.

The experiments in ontology mappings have been performed by adopting knowledge based approaches, adopting already existing ontologies available on the Semantic Web. The technology developed is highly general and domain independent, since it relies on SWOOGLE and WATSON as general purpose and open domain search engines for the semantic WEB. Therefore, it can be applied to any ontology, language and domain, provided that the SW contains enough knowledge related to the domain.

Results obtained from our experiments were controversial, since the accuracy of the method varies from 0,28 for the AGROVOC-ASFA matching to 0,70 for the AGROVOC-NALT matching. A possible explanation for this large difference could be the weakness in the semantic web to fit the relations required for the first case. Anyhow, since the semantic web is growing very fast, improvements could be expected in the next years, when much more knowledge will be available. Given the disparity in the results obtained, we do not recommend the integration of this methodology as it is, since the effort required to validate non accurate results would be higher than the effort required to perform the task manually. We also recommend further investigation into ontology mapping in order to produce a methodology capable of providing much more accurate results, for example by detecting incorrect mappings automatically on the basis of distributional similarity in corpora, and only then proceeding further in the integration phase.

Ontology Learning Techniques not covered in this Deliverable

In this deliverable, we suggested solutions to fundamental problems in ontology learning and alignment that are considered most relevant for the WP7 case study. The proposed frameworks and techniques can be used, e.g., to acquire new terms, class instantiations or object properties. For further refinement and evaluation of the fishery ontologies and their respective alignments we additionally recommend an investigation into the usefulness of tools for acquiring more expressive ontological constructs, experimented into different WPs of the NeOn project. RELExO (Völker and Rudolph, 2008), for example, could be applied to assure both quality and logical completeness of the ontologies, whereas tools such as LeDA (Völker et. al, 2007) seem to be a promising way to facilitate the detection of modeling errors by automatically generated disjointness axioms. If necessary, inconsistencies introduced in this process of ontology enrichment will have to be resolved by approaches to handling inconsistency and incoherence as described in D1.2.1. In a detailed evaluation based on case study data we showed that provenance information produced in the course of ontology learning is a valuable help in debugging learned ontologies (see D1.2.2).

6 References

Z. Aleksovski, M. Klein, W. ten Katen, and F. van Harmelen. Matching Unstructured Vocabularies using a Background Ontology. In Proc. of EKAW, 2006.

Y. Arens, C.Y. Chee, C.N. Hsu, and C.A. Knoblock. Retrieving And Integrating Data From Multiple Information Sources. International Journal of Cooperative Information Systems, 2(2):127 – 158, June 1993.

Luisa Bentivogli, Pamela Forner, and Emanuele Pianta. Evaluating cross-language annotation transfer in the multisemcor corpus. In COLING '04: Proceedings of the 20th international conference on Computational Linguistics, page 364, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

P. Bouquet, L. Serafini, and S. Zanobini. Peer-to-Peer Semantic Coordination. Journal of Web Semantics, 2(1), 2005.

P. Buitelaar, P. Cimiano, and B. Magnini. 2005. Ontology learning from texts: methods, evaluation and applications. IOS Press.2002.

Buitelaar P., Olejnik D., Sintek M. A Protege Plug-In for Ontology Extraction from Text Based on Linguisitc Analysis. In: Proceedings of the European Semantic Web Symposium ESWS-2004, Greece, May 2004

Ciaramita, M.; Altun, Y. Broad-Coverage Sense Disambiguation and Information Extraction with a Supersense Sequence Tagger, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2006

P. Cimiano, S. Handschuh, and S. Staab. Towards the Self-Annotating Web. In Proc. of WWW, 2004.

Philipp Cimiano, Johanna Völker. Text2Onto - A Framework for Ontology Learning and Data-driven Change Discovery. In Andres Montoyo, Rafael Munoz, Elisabeth Metais, Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB), volume 3513 of Lecture Notes in Computer Science, pp. 227-238. Springer, Alicante, Spain, June 2005.

Cunningham, H. Maynard, D. Bontcheva, K. and Tablan, V. (2002), GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

Cohn, D. A., Ghahramani, Z., and Jordan, M. I. (1995). Active learning with statistical models. In Tesauro, G., Touretzky, D., and Leen, T., editors, Advances in Neural Information Processing Systems, volume 7, pages 705–712. The MIT Press.

C. Collet, M.N. Huhns, and W. Shen. Resource Integration Using a Large Knowledge Base in Carnot. IEEE Computer, 24(12):55 – 62, Dec 1991.

Daelemans, W., and Van den Bosch, A. (2005), Memory-Based Language Processing, Cambridge, UK: Cambridge University Press.

M. d'Aquin, M. Sabou, M. Dzbor, C. Baldassarre, L. Gridinoc, S. Angeletou, and E. Motta. WATSON: A Gateway for the Semantic Web. In Proc. of ESWC, Poster Session, 2007.

S. Deerwester and S. Dumais and G. Furnas and T. Landauer and R. Harshman *Indexing by Latent Semantic Analysis*, Journal of the American Society of Information Science, 1990.

F. de Saussure. 1922. Cours de linguistique générale. Payot, Paris. P. Pantel and D. Lin. 2002.

U. Eco. 1979. Lector in fabula. Bompiani.

M. Ehrig and Y. Sure. FOAM - Framework for Ontology Alignment and Mapping; Results of the Ontology Alignment Initiative. In Proc. of the Workshop on Integrating Ontologies. 2005.

J. Euzenat, M. Mochol, P. Shvaiko, H. Stuckenschmidt, O. Svab, V. Svatek, W. R. van Hage, and M. Yatskevich. Results of the Ontology Alignment Evaluation Initiative 2006.

R. Girju, A. Badulescu, and D. Moldovan. 2006. Learning semantic constraints for the automatic discovery of part-whole relations. In Proceedings of HLT/NAACL- 03, pages 80–87, Edmonton, Canada, July.

F. Giunchiglia, P. Shvaiko, and M. Yatskevich:. Semantic Schema Matching. In Proc. of CoopIS'05, volume 3760 of LNCS, pages 347 – 360, 2005.

F. Giunchiglia, P. Shvaiko, and M. Yatskevich. Discovering Missing Background Knowledge in Ontology Matching. In Proc. of ECAI, 2006.

Alfio Massimiliano Gliozzo *Semantic Domains in Computational Linguistics*, PhD thesis University of Trento, 2005

Alfio Massimiliano Gliozzo,Marco Pennacchiotti and Patrick Pantel. 2007. "The Domain Restriction Hypothesis: Relating Term Similarity and Semantic Consistency". In Proceedings of North American Association for Computational Linguistics / Human Language Technology (NAACL HLT 07). pp. 131-138. Rochester, NY.

Gruber T. Towards principles for the design of ontologies used for knowledge sharing. Int. Journal of Human and Computer Studies 43(5/6), 1994, 907-928.

M.A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th International Conference on Computational Linguistics. Nantes, France.

W. Hu, G. Cheng, D. Zheng, X. Zhong, and Y. Qu. The Results of Falcon-AO in the OAEI 2006 Campaign.

M.N. Huhns, N. Jacobs, T. Ksiezyk, W. Shen, M.P. Singh, and P.E. Cannata. Integrating enterprise information models in Carnot. In Proc. of Intel ligent and Cooperative Information Systems, 1993.

N. Jian, W. Hu, C. Cheng, and Y. Qu. Falcon-AO: Aligning Ontologies with Falcon. In Proc. of the Workshop on Integrating Ontologies. 2005.

Y. Kalfoglou and M. Schorlemmer. Ontology Mapping: The State of the Art. The Knowledge Engineering Review, 18(1):1–31, 2003.

S. Deerwester and S. Dumais and G. Furnas and T. Landauer and R. Harshman *Indexing by Latent Semantic Analysis* Journal of the American Society of Information Science 1990

Y. Li, J. Li, D. Zhang, and J. Tang. Result of Ontology Alignment with RiMOM at OAEI06.

Alexander Mädche, Steffen Staab. Ontology Learning. In Steffen Staab and Rudi Studer, Handbook on Ontologies, International Handbooks on Information Systems, pp. 173-190, Springer, 2004.

B. Magnini and G. Cavaglià *Integrating Subject Field Codes into WordNet* Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation 2000

B. Magnini, B., Negri, M., Prevete, R.,and Tanev, H. (2002), A WordNet-based approach to named entities recognition. In Proceedings of SemaNet'02: Building and Using Semantic Networks, pages 38-44, Taipei, Taiwan, August 2002.

McDonald, D. (1996), Internal and external evidence in the identification and semantic categorization of proper names. In: Boguraev, I., Pustejovsky, J. (eds.): Corpus Processing for Lexical Acquisition, Chapter 2. The MIT Press, Cambridge, MA (1996).

Maedche A. Ontology Learning for the Semantic Web. The Kluwer International Series in Engineering and Computer Science, Volume 665, 2003.

M. Mao and Y. Peng. PRIOR System: Results for OAEI 2006.

S. Massmann, D. Engmann, and E. Rahm. COMA++: Results for the Ontology Alignment Contest OAEI 2006.

Mikheev, A. and Grover, C. and Moens, M. (1999), Tools and Architecture for Named Entity Recognition, Journal of Markup Languages: Theory and Practice, vol.1,no. 3, 1999, pp.89-113

Mikheev, A., Moens, M., Grover, C. (1999), Named Entity recognition without gazetteers. Proceedings of EACL-99, Bergen, Norway (1999).

R. Navigli, P. Velardi. Learning Domain Ontologies from Document Warehouses and Dedicated Websites. Computational Linguistics, 30(2), MIT Press, 2004, pp. 151-179.

Navigli R., Velardi P and Gangemi A. Ontology Learning and its application to automated terminology translation. IEEE Intelligent Systems, vol. 18:1, January/February 2003.

NeOn deliverable D3.4.2.

N.F. Noy. Semantic Integration: a Survey Of Ontology-Based Approaches. SIGMOD Record, 33(4):65–70, December 2004.

P. Pantel and M. Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In ACL-COLING-06, pages 113–120, Sydney, Australia.

M. Pasca and S. Harabagiu. 2001. The informative role of wordnet in open-domain question answering. In Proceedings of NAACL-01 Workshop on WordNet and Other Lexical Resources, pages 138–143, Pittsburgh, PA.

M. Pazienza, M. Pennacchiotti, and F. Zanzotto. Terminology extraction: an analysis of linguistic and statistical approaches. In S.Sirmakessis, editor, Knowledge Mining, volume 185. Springer Verlag, 2005.

Davide Picca, Alfio Massimiliano Gliozzo and Massimiliano Ciaramita. Semantic Domains and Supersense Tagging for Domain-Specific Ontology Learning, in Prodeedings of RIAO'2007, Pittsburgh.

Davide Picca, Alfio Massimiliano Gliozzo and Massimiliano Ciaramita, Italian Porting of Supersense Tagger, forthcoming

E. Rahm and P. Bernstein. A Survey of Approaches to Automatic Schema Matching. The VLDB Journal, 10:334–350, 2001.

M. Sabou, M. d'Aquin, and E. Motta. Using the Semantic Web as Background Knowledge for Ontology Mapping.

R. Snow, D. Jurafsky, and A.Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In Proceedings of the ACL/COLING-06, pages 801–808, Sydney, Australia.

P. Shvaiko and J. Euzenat. A Survey of Schema-based Matching Approaches. Journal on Data Semantics, IV, 2005.

P. Shvaiko, J. Euzenat, N. Noy, H. Stuckenschmidt, R. Benjamins, and M. Uschold, editors. Proc. of the ISWC Ontology Matching WS, 2006.

L.M. Stephens, A.K. Gangam, and M.N. Huhns. Constructing Consensus Ontologies for the Semantic Web: A Conceptual Approach. World Wide Web Journal, 7(4):421–442, December 2004.

H. Stuckenschmidt, F. van Harmelen, L. Serafini, P. Bouquet, and F. Giunchiglia. Using C-OWL for the Alignment and Merging of Medical Ontologies. In Proc. of the First Int. WS. on Formal Biomedical K. R. (KRMed), 2004.

W. van Hage, S. Katrenko, and G. Schreiber. A Method to Combine Linguistic Ontology-Mapping Techniques. In Proc. of ISWC, 2005.

Johanna Völker and Sebastian Rudolph. Lexico-Logical Acquisition of OWL DL Axioms - An Integrated Approach to Ontology Refinement. In Raoul Medina and Sergei Obiedkov, Proceedings of the 6th International Conference on Formal Concept Analysis (ICFCA'08). Springer, February 2008. to appear

Johanna Völker, Denny Vrandecic, York Sure, Andreas Hotho. Learning Disjointness. In Enrico Franconi, Michael Kifer, Wolfgang May, Proceedings of the 4th European Semantic Web Conference (ESWC'07), volume volume 4519 of Lecture Notes in Computer Science, pp. 175-189. Springer, June 2007.

H. Wache, T. Voegele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Huebner. Ontology-Based Integration of Information. A Survey of Existing Approaches. In Proc. of the WS on Ontologies and Information Sharing, 2001.

Wakao, T., Gaizauskas, R., Wilks, Y. (1996), Evaluation of an Algorithm for the Recognition and Classi_cation of Proper Names. Proceedings of the 16th Conference on Computational Linguistics (COLING '96), Copenhagen, Denmark (1996).